

Feb 02, 2026

Version 6

GenFS Metadata Cleanup Challenge protocol V.6

DOI

<https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v6>



Ruth Timme¹, Martin Shumway², Maria Balkey³, Tina Pfefer⁴

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

GenomeTrakr

Tech. support email: 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/v6>

Protocol Citation: Ruth Timme, Martin Shumway, Maria Balkey, Tina Pfefer 2026. GenFS Metadata Cleanup Challenge protocol. protocols.io <https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v6> Version created by **Maria Balkey**



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: January 22, 2026

Last Modified: February 02, 2026

Protocol Integer ID: 239147

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

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 2025 Metadata Cleanup Challenge
- Preparing the update file
- Making corrections and updates within the file
- Submitting the finalized template

Version History:

V6: Remove video instructions, adding section for purpose of sampling, updating picklists for mandatory OHE attributes, replacing Example Template excel file, and text edits to improve clarity.

V5: addition of picklists for mandatory OHE attributes

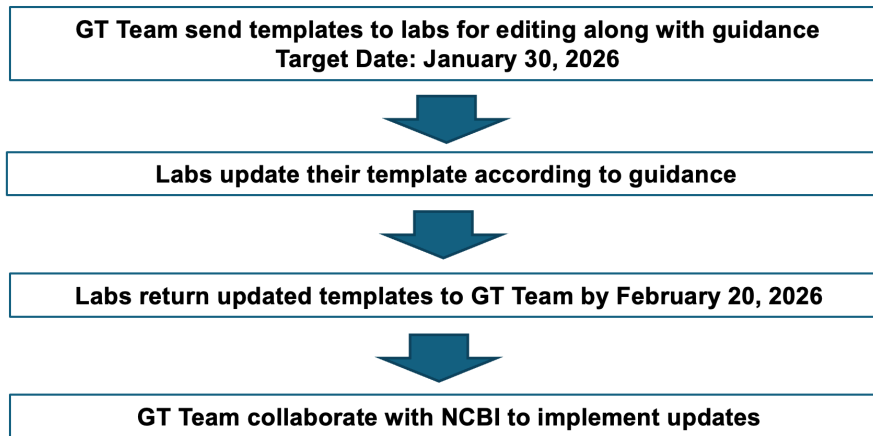
V4: addition of a training video

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 2025 Metadata Cleanup Challenge

1

Metadata Cleanup Project Workflow



- 2 This exercise is intended to support review, standardization, and curation of core metadata required under the **NCBI One Health Enteric package**.

Timeframe

- The primary focus is on submissions made between **September 2024 and August 2025**.
- Laboratories may include submissions outside this timeframe if doing so improves metadata completeness or consistency.

Review and curate the following **core One Health Enteric attributes**:

- source_type
- collected_by
- sequenced_by
- project_name
- purpose_of_sampling
- isolation_source
- host (*human and animal isolates only*)



- food_origin (*food isolates only*)

Guidance for preparing the update file

3 For labs whose BioProjects are linked to the GenomeTrakr Umbrella BioProject:

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

Access your lab-specific metadata template here:

DOCS → *metadata hackathon* → *2025 Metadata Template Pickup*

DOCS: metadata hackathon/2025 Metadata Template Pickup

4 For labs whose BioProject is *not* linked to the GenomeTrakr umbrella

If your lab does not have a BioProject linked to PRJNA593772, you can generate your own metadata template using the NCBI SRA Run Selector as described below.

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

4.2 In the search field, enter your BioProject accession(s), separated by commas if more than one

- Example:

PRJNA593772, PRJNA593774



SRA Run Selector 🔍 ? ⚙️ 🔗

Accession PRJNA692474 🔍 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-east1, ncbi.public, s3.us-east-1
LibraryLayout	PAIRED
LibrarySource	GENOMIC
Platform	ILLUMINA
version	1

Select

	Runs	Bytes	Bases	Download
Total	241	66.33 Gb	136.36 Gb	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List

4.3 After the results load, locate the **Download** panel and click the **"Metadata"** button.

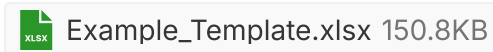
Preparing the metadata file

- The downloaded file will be a **comma-delimited txt file** containing both **BioSample** and **SRA** metadata.

Open the file in **Excel** and edit as follows:

- Remove all metadata columns that are **not relevant** for this exercise.
- Move the **BioSample accession** column to be the **first column**.
- Save the file and proceed with metadata updates.

For guidance on which columns to retain, refer to the example template:



If you identify any isolates in your template that should instead be submitted under another organization's program (e.g., PulseNet), please notify the GenomeTrakr team: @genometrakr@fda.hhs.gov.

5 Prepare the update file:

File organization



- Submit **one bulk update file** per:
 - laboratory, **or**
 - laboratory + organism, **or**
 - another logical aggregation agreed upon in advance.

File Structure:

Rows

- The **first row** must contain the **column headers**.
- Each subsequent row must represent **exactly one BioSample** and its associated updatable attributes.
- All rows must contain the **same number of columns**.

Columns:

- Column headers must correspond to **attribute names** included in the **One Health Enteric package** (either full attribute names or harmonized names).
- The **first column** must contain the **BioSample accession (e.g. SAMN123456789)**

File format

- The update file must be a **tab-delimited text file** with a **.tsv** or **.txt** file extension.
- If you use Microsoft Excel for editing, **export the final file** in tab-delimited format.
- Verify that **date fields are correctly exported** and retain the intended format.

FYI:

Only include columns for which **at least one BioSample** has a value to be updated.

For any given BioSample:

- **Blank field in update file (with existing value in NCBI)** → existing value will be *replaced with NULL*.
- **Field already correct in NCBI** → populate the field with its *current value* to prevent changes.
- **Field not currently present in NCBI** → value will be *added*.
- **Blank field for a non-existent attribute** → attribute 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 updating **MANDATORY** fields

- 6 In your Excel template, ensure there's an entry in these columns for every record (rpw). If information is missing, choose one of the null terms, "Not Applicable, Not Collected, Not



Provided, Missing, or Restricted Access".

7 **source_type**

Review the metadata file and confirm that every BioSample record includes a **source_type** value. Update or correct entries as needed.

Allowed values (controlled vocabulary)

source_type is a required, controlled attribute. Only the values listed below are permitted — no substitutions, abbreviations, or alternate spellings are allowed

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

Entries that do not exactly match one of these terms will **fail validation**.

8 **project_name**

Review the entries in this column and **populate this field for every record**. Correct any existing values as needed.

For U.S. pathogen surveillance submissions, select the coordinating program or network that best represents the isolate.

Commonly used project_name values:

Select one or more of the following, as applicable:

- GenomeTrakr
- GenomeTrakr; LFFM-FY1
- GenomeTrakr; LFFM-FY2
- GenomeTrakr; LFFM-FY3
- GenomeTrakr; LFFM-FY4
- GenomeTrakr; LFFM-FY5
- GenomeTrakr; LFFM-FY6
- NARMS
- NARMS Cecal
- NARMS Retail Meat



- PulseNet
- USDA-FSIS
- Vet-LIRN
- NAHLN

Multiple project affiliations

- If an isolate belongs to more than one project or network, include all applicable terms.
- Separate multiple values using a semicolon followed by a space, example:
GenomeTrakr; USDA-ARS

Adding new project_name terms

- If your isolate is associated with a project or network *not* listed above, you may:
- Enter a new term directly into the update template, and
- Request that the term be added to the official picklist by emailing:
genometrakr@fda.hhs.gov
- Newly requested terms will be reviewed and added in the **next picklist update**, as appropriate.

9 purpose_of_sampling

Review the entries in the **purpose_of_sampling** column and populate this field for every record. Correct any incorrect or inconsistent values.

- Use **only values from the One Health Enteric package picklist**.
- Enter **one value per record**.
- Do not use free text, abbreviations, or alternate spellings.

Allowed values (controlled vocabulary)

targeted surveillance/monitoring

baseline surveillance/monitoring

diagnostic testing

research

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

Guidance:

targeted surveillance/monitoring: Sampling conducted in response to a specific public health, regulatory, or food safety event (e.g., outbreak investigation, compliance or enforcement action, contamination event, or other defined issue).



baseline surveillance/monitoring: Sampling conducted as part of routine, ongoing surveillance activities without a triggering event (e.g., routine PulseNet submissions, routine GenomeTrakr inspections, recurring food sampling plans, routine environmental or agricultural water monitoring).

diagnostic testing: Sampling performed primarily for clinical or laboratory diagnosis, not for surveillance.

research: Sampling conducted for research or method development purposes, not for routine surveillance or regulatory response.

10 collected_by and sequenced_by

Review the values in both the collected_by and sequenced_by columns and ensure that every record is populated. Correct any errors or inconsistencies.

Standardization requirements

- Use standardized laboratory names consistently across all records.
- For sequenced_by select the laboratory name exactly as listed in the current One Health Enteric package picklist.
- Do not use abbreviations, alternate spellings, or free-text variations.

Updating laboratory names

If your laboratory name is missing, incorrect, or has changed:

- Email the corrected or preferred laboratory name to: **genometrakr@fda.hhs.gov**
- The GenomeTrakr team will update the picklist to reflect the change in a future release.

11

Guidance for Conditionally Required fields

12 Conditionally required attributes - Overview

For this exercise, two attributes are conditionally mandatory. These fields must be populated when specific sample types are present, as determined by **source_type**.

Conditionally required fields

- **host:** Required for isolates derived from human or animal samples.
- **food_origin:** Required for isolates derived from food products or other commercial products sampled for pathogens, including (but not limited to): foods, medical products, cosmetics, tattoo ink, etc.



How to identify applicable records

Use the column to filter your dataset and identify records requiring these fields:

1. Filter for **human** and **animal** records → populate **host**
2. Filter for **food** records → populate **food_origin**

Records with other **source_type** values do not require these attributes unless explicitly instructed

13 **host**

(Human and animal isolates only)

The **host** field is required for all isolates derived from human or animal samples.

How to populate:

- Filter for **source_type** = **human** or **animal** in Excel.
- Do **not** populate **host** for non-host-associated sample types.

How to determine the host value

- Use the **isolation_source** column to inform the appropriate host entry.
- Enter the **scientific (binomial) name** whenever possible, examples: *Homo sapiens*, *Bos taurus*
- If the scientific name is unknown, enter a common name recognized by the **NCBI taxonomy database** examples: porcine, bovine

Unknown host

If the host cannot be determined, enter **one** of the following NCBI-approved null values:

- Not Applicable
- Not Collected
- Not Provided
- Missing
- Restricted Access

Guidance for updating MANDATORY fields

14 **food_origin**

(Food or other product isolates only)

The **food_origin** field is required for isolates derived from food products or other commercial products sampled for pathogens.

Identify applicable records:



- Sort or filter the dataset using the **source_type** column.
- Select records where **source_type** = food.

Understanding location fields

Food and product isolates use two separate geographic attributes:

- **geo_loc_name** – Location where the sample was physically collected (e.g., U.S. state, port of entry, retail location)
- **food_origin** – Geographic origin of the food or product sampled (e.g., country or state where the product was produced)

How to populate food_origin

Review the values in **geo_loc_name**.

Apply the following rules:

If **geo_loc_name** currently reflects the origin of the food product (e.g., "India" for an imported food sampled in the U.S.):

- Move that geographic value to **food_origin**
- Update **geo_loc_name** to reflect the actual collection location, if known

If **geo_loc_name** reflects the sample collection location (e.g., port of entry, U.S. state, retail site):

- Leave **geo_loc_name** unchanged
- Populate **food_origin** with the country or state of origin, if known

Unknown food origin: If the origin of the food or product cannot be determined, enter one of the following NCBI-approved null values:

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

15 **isolation_source**

1. Review entries and clean up spelling errors and/or non-standard terms

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

OPTIONAL: bring your records up to OHE standards

- 16 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 sub-packages.

16.1 **animal samples**, `source_type = animal`

▪ ***animal_env* picklist:**

veterinary facility or diagnostic laboratory
 private household [ENVO:01000418]
 animal breeding facility
 animal boarding facility [ENVO:00003040]
 animal exhibition site
 animal market or collection point



animal import/export quarantine facility
abattoir [ENVO:01000925]
food animal production site
zoo, wildlife refuge, or private animal collection [ENVO:00010625]
natural environment [ENVO:01000951]
fish hatchery [ENVO:00000295]
poultry hatchery [ENVO:01001874]
smallholder farm
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

16.2 **food product samples**, source_type = food

▪ ***intended_consumer* picklist:**

human as food consumer
animal as food consumer
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

▪ ***food_processing_method* picklist**

food (raw) [FOODON:03311126]
food (heat treated) [FOODON:03316043]
food (cooked) [FOODON:00001181]
food (precooked, frozen) [FOODON:03305323]
food (pasteurized) [FOODON:00002654]
aseptic filling and sealing of food [FOODON:03470126]
ultra pasteurization (UP) by heat [FOODON:03470148]
food (preserved) [FOODON:00002158]
food (pickled) [FOODON:00001079]
food (freeze-dried) [FOODON:03301752]
food (acidified) [FOODON:03301625]
food (low acid, canned) [FOODON:03301624]
food (canned) [FOODON:00002418]
food (refrigerated) [FOODON:03470172]
food (frozen) [FOODON:03302148]
food (fresh frozen) [FOODON:00004731]
food (smoked) [FOODON:03310311]



food (dehydrated) [FOODON:00002643]
food (fermented) [FOODON:00001258]
food (irradiated) [PATO:0001744]
food (salted) [FOODON:03460173]
preservation by storage in modified atmosphere [FOODON:0347011]
food (breaded) [FOODON:00002661]
food (cleaned) [FOODON:00002708]
food (comminuted) [FOODON:00002754]
food (cut) [FOODON:00004291]
food (filled) [FOODON:00002644]
food (ground) [FOODON:00002713]
food (juiced) [FOODON:00003499]
food (milled) [FOODON:00002649]
food (packaged) [FOODON:00002739]
food (peeled) [FOODON:00002655]
food (puffed) [FOODON:00002656]
food (rehydrated) [FOODON:00002755]
food (textured) [FOODON:00002658]
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

16.3 **facility inspection samples**, source_type = environmental

▪ ***facility_type* picklist:**

ambient storage
caterer-catering point
distribution
farm
frozen storage
importer-broker
interstate conveyance
labeler-relabeler
packaging
process/manufacturing
refrigerated storage
storage
Not Applicable
Not Collected
Not Provided
Missing

Restricted Access

▪ ***food_type_processed*** picklist:

Animal feed
Baby Food products
Bakery Products, Doughs, Bakery Mixes and Icings
Candy, candy specialties, chewing gum
Cereal preparations and Breakfast foods
Cheese and Cheese Products
Chocolate, cocoa products, cocoa beans
Coffee and tea
Dietary supplements
Drinks, soft drinks, and waters
Edible insects and insect-derived foods
Egg and Egg Products
Fishery/Seafood Products
Fruit and Fruit Products
Ice Cream and related products
Macaroni and noodle products
Meats, meat products, and poultry
Medicated animal feeds
Milk Butter and Dried Milk Products
Miscellaneous food related items
Multiple Food Dinners, Gravies, Sauces and Specialties
Nuts and Edible seeds
Pet and laboratory animal food
Pet food and treats
Prepared salad products
Snack food items
Soups
Spices, flavors, and salts
Vegetable oils
Vegetable protein products
Vegetables and Vegetable Products
Whole grains, milled grain products, and starches
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

16.4 **farm/environment samples**, source_type = environmental

**■ env_broad_scale picklist:**

agricultural ecosystem [ENVO:00000077]

aquatic ecosystem [ENVO:01001787]

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

■ env_local_scale picklist:

farm [ENVO:00000078]

produce farm

mixed-use farm

livestock operation

pasture [ENVO:00000266]

fish farm [ENVO:00000294]

feedlot [ENVO:01000627]

under glass/protected plant cultivation [FOODON:03530211]

aquaculture

indoor rearing structure

outdoor rearing structure

stream [ENVO:00000023]

river [ENVO:00000022]

lake [ENVO:00000020]

pond [ENVO:00000033]

canal [ENVO:00000014]

plant body [PO:0009011]

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

■ env_medium picklist:

soil [ENVO:00001998]

hay for animal feed [FOODON:03301763]

hay [baled] [FOODON:03309364]

straw [FOODON:03309894]

animal litter [ENVO:00002191]

animal manure [ENVO:00003031]

air [ENVO:00002005]

saline water [ENVO:00002010]

freshwater [ENVO:00002011]



sediment [ENVO:00002007]
wastewater [ENVO:00002001]
sewage [ENVO:00002018]
plant root, tuber or bulb [FOODON:03420238]
plant part above surface [FOODON:03420144]
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

Finalize your update file(s) and submit to protocols.io for NCBI processing

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

17.1 Bring your template to the **February 11th, 2026** GenomeTrakr metadata office hours. We will provide assistance to labs that might need input from GenomeTrakr to complete metadata templates.

17.2 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

17.3 Save as a tab-delimited .tsv file and transfer the .tsv file to protocols.io:
DOCS: metadata hackathon/2025 Metadata Template Submission

17.4 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.

