Designed by Charlie Barclay, Emma Griffiths and Rhiannon Cameron as part of PHA4GECon 2025 pre-conference workshops

PHA4GECon 2025: Ontology Term Curation: Cleaning and Standardising Contextual Data

Objective

Learn to curate unharmonised and inconsistent (“dirty”) contextual data using ontology look-up tools, then apply harmonised terms to complete a schema.

Resources

For this exercise you will need to download the following files:

- PHA4GECon: Standardise your dataset **a mock dataset of un-standardised data*
- PHA4GECon: SARS-CoV-2 schema

Instructions

Part 1: Tidying your data

Scenario: You have been tasked with harmonising data from two labs into a single dataset. You have been provided with a minimal set of completed fields in a single dataset. The goal is to clean this data, such that everyone is using the same controlled vocabulary to represent the same thing within a field.

1. Open the workbook ‘standardise-your-dataset.xlsx’ under workshop/PHA4GECon-2025/2_cleaning-data and locate the ‘Original Data’ tab. *Question: What do you notice about this data?*
2. Pull out all **unique terms** (this has been done for you in the ‘Part 1’ tab)
3. Pick two groups (e.g. ‘organism’, ‘host’, or ‘anatomical parts’ etc)
4. Evaluate for duplication, inconsistent formatting, or ambiguity. Choose one ‘harmonised’ label e.g for ‘Nasopharynx’ and ‘NP’ I might select ‘Nasopharynx (NP)’.

5. Produce a short list of cleaned, harmonised English labels—one per concept—to carry forward.

Tips: *Think conceptually, not literally*: Focus on what the term means, not just how it's written. "Plant" in a wastewater context = facility, not flora.

Group before cleaning: Gather all unique values for each field first—seeing variants side-by-side helps spot duplicates and inconsistencies.

Handle missing data early: Collect all "unknown", "N/A", or blank entries and map them to an agreed missing-value term (e.g., missing [GENEPIO:0001618]).

Part 2: Finding ontology matches

- Using your cleaned list, search for corresponding ontology terms that best match each harmonised label using
    - The EBI Ontology Lookup Service (https://www.ebi.ac.uk/ols/index)
    - OntoBee (https://ontobee.org)
- Evaluate search results and choose the most precise match. Consider whether the match is appropriate (too broad, too narrow, wrong domain?).
- Record for each:
    - Harmonized Label
    - Ontology Name
    - Ontology ID (e.g. ENVO:00002272)
      Definition (if available)
- If a good match cannot be found, note **"New term request needed"**.

Tips: *Check the definition, not just the label:* Many terms look similar—read the ontology definition to confirm the meaning fits your data.

Prefer OBO Foundry sources such as ENVO (environmental), OBI (process/sample type), or UBERON (anatomy).

Reflection

Challenges	Details
