Improving Openness And Interoperability In Public Health Bioinformatics

A Global Coalition.

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

PHA4GECon 2025: Data Curation Exercise

Purpose

This exercise demonstrates how data specification are operationalised through tooling, in this instance the DataHarmonizer - a spreadsheet style data validation application.

Using the PHA4GE SARS-CoV-2 Contextual Data Template, you will curate one record from a clinical scenario, validate it, and discuss how tooling can improve data quality and usability.

Getting Started with the DataHarmonizer

Go to the https://github.com/cidgoh/pathogen-genomics-package and go to the latest releases to download the lates package. Unzip the package and run by loading index.html or main.html.

The default template loaded is the "CanCOGeN Covid-19" template. To change the spreadsheet template, select the white text box to the right of Template, it always contains the name of the template currently active, or navigated to File followed by Change Template. An in-app window will appear that allows you to select from the available templates in the drop-down menu. After selecting the desired template, click Open to activate the template.

Scenario

DetailsThe BCCDC Public Health Laboratory obtained a nasopharyngeal swab for diagnostic testing (sample ID Bc-12345-ab) on March 1 2023 from a symptomatic, 44 year old female that had been hospitalized in the ICU. The individual had been exhibiting a cough, fever, muscle weakness, as well as other symptoms of Acute Respiratory Distress Syndrome. The individual recently travelled to the United States on holiday and returned on Feb 19 2023. The sample was flagged for sequencing as part of the lab's International travel surveillance program. The sample was sequenced on March 7 2023 using an Illumina MiSeq instrument. The raw data was processed using ncov-tools 2.3.1 as part of their bioinformatics protocol (https://github.com/jts/ncov2019-artic-nf/blob/master/README.md) and dehosted using BWA (version 0.7.17). The consensus sequence was generated using iVar 2.3.1. The sequence was uploaded to GISAID and assigned the accession number EPI_ISL_436489. Drs Tejinder Singh, Fei Hu and Joe Blogs helped to generate the sequence.

Instructions

- Review the scenario and pull out the key pieces of information
- Populate the template, paying care to the required fields
    - If you are uncertain what a field is meant to capture, hover ove the field and a pop up will appear with the appropriate definition.
- Click Validate on the toolbar.
- Review any warnings or errors.
- Correct missing values or formatting issues:
    - All dates must use YYYY-MM-DD.
    - Required fields (yellow) cannot be blank.
    - Ontology terms should display as Label [ID].