Infectious Disease Genomic Epidemiology - Data curation and sharing lab instructions

April 2023

The IDGE Data curation and sharing lab is designed to give participants hands-on experience using curation tools to structure, standardize, and transform genomics contextual data.

The lab experience is divided into four parts that consist of:

1) A **demo** of the DataHarmonizer.
2) A **data curation exercise** where participants practice interpreting and structuring public health contextual data using the DataHarmonizer. Curated data will then be transformed into different repository submission formats for comparison.
3) A **data standardization exercise** where participants will practice using an online look-up service to identify standardized terms. These skills can be applied later to the participants' own data.
4) A **review of the World Health Organization's recommendations** for pathogen genomics data sharing.

Learning Objectives:

1. Understand the importance of data curation for improving and ensuring contextual data quality.
2. Understand the ways that data standards can facilitate data quality.
3. Be able to describe which data types are important for pathogen genomic surveillance.
4. Be able to describe the differences and similarities in public repository submission requirements.
5. Be familiar with contextual data curation and standardization tools.
6. Be aware of global expectations for genomic surveillance data sharing.
7. Be able to discuss the benefits and risks of pathogen genomics data sharing.

Part 1 - DataHarmonizer demonstration.

The instructor will provide a brief overview of the DataHarmonizer tool, with examples of how to enter, validate and transform data for repository submission.

Part 2 - Curating contextual data using the DataHarmonizer.

Participants will practice downloading and using the DataHarmonizer application. Participants will be provided with a set of scenarios of mock public health pandemic contextual data and will be asked to enter pertinent details into the application. A curation standard operating procedure (SOP) will also be distributed to provide guidance for interpreting the scenarios and to provide additional ethical, practical, and privacy considerations when curating.

Instructions:

1. Download a copy of the DataHarmonizer.

- Download the zip file ("Source code (zip)") containing The "Pathogen Genomics Package" version of the DataHarmonizer application from the following link: https://github.com/cidgoh/pathogen-genomics-package/releases/tag/PHPv2.0.3.
- Extract the zip file's contents, and navigate into the extracted folder. Open "**main.html"**. The validator application will open in your default browser[1].
- The CanCOGeN[2] template will open as a default setting - you are ready to curate!

2. Review the public health contextual data scenarios and the curation SOP.

- Review the DataHarmonizer Curation SOP: CBW Metadata Curation SOP_1.0
- Read over the four different scenarios and identify the pieces of information to be entered into the app based on the instruction provided in the demo.
- Scenarios doc: CBW2023 Contextual Data Curation Scenarios
- Consult the curation SOP for examples and additional guidance.

3. Enter information into the DataHarmonizer app and validate the data.

- Open the DataHarmonizer, beginning with the empty CanCOGeN template - filling in scenario 1 during the demo (or "**Open**" this template file which has the scenario 1 demo already filled in).
- Use the fields and dropdown menus to enter data into the app. Standardize as you go by checking the headers for formatting recommendations, checking the curation SOP, and selecting controlled vocabulary from picklists when available.
- In the interest of time, try to input data for every field from all the case scenarios. **Hint:** you can copy and paste in the DataHarmonizer just like in any other spreadsheet.
- When complete, click "**Validate**" in the menu to check for errors and missing information.
- Make any corrections necessary.

---

[1] The DataHarmonizer is compatible with Chrome (49+), Firefox (34+), and Edge (12+)
[2] Canadian COVID-19 Genomics Network (CanCOGeN) - genomecanada.ca/challenge-areas/cancogen

4. Export the data in GISAID[3] and NCBI BioSample[4] formats for comparison.

- Once the data has been validated with no errors, export the data in submission-ready formats by clicking "**File**" followed by "**Export to**….
- Select "**GISAID**" from the "**Format**" menu and name the file for download.
- Repeat the export process, but this time select "**BioSample**" from the "**Format**" menu and name the file for download.
- Open the two files and explore the similarities and differences. Consider the following:
    - How are the formats different?
    - What information is the same?

5. Participate in the group discussion about the activity.

- Reflect on your experience curating, validating and transforming contextual data using the DataHarmonizer.
    - Was it easy to use?
    - Were all the fields and terms you needed there?
    - Were the definitions and examples in the reference guide helpful?

Further learning:

Protocols for setting up accounts and submitting to GISAID[3], NCBI[4] and ENA[5] are available on protocols.io (https://www.protocols.io/workspaces/pha4ge). The protocols were developed by the Public Health Alliance for Genomic Epidemiology (PHA4GE) - an international community of scientists from public health, industry and academia focused on improving the reproducibility, interoperability, portability and openness of public health bioinformatic software, skills, tools and data.

Learn more about the development and international use of the SARS-CoV-2 contextual data specification by reading PHA4GE's paper "Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package".

Part 3 - Using ontology look-up services to standardize data

Participants will learn to use the European Bioinformatics Institute's ontology look-up service (EBI-OLS) to identify standardized terms. These skills can be applied to standardizing data in the absence of consensus data standards (i.e. when a template is unavailable for particular data types) or for identifying additional standardized terms which can be added to existing specifications.

---

[3] Global Initiative on Sharing Avian Influenza Data (GISAID) - www.gisaid.org
[4] National Center for Biotechnology Information (NCBI) BioSample database - www.ncbi.nlm.nih.gov/biosample
[5] European Nucleotide Archive (ENA) - www.ebi.ac.uk/ena

Instructions:

1. Watch the EBI OLS demo.

- Watch the brief demo by the instructor and note the guidance on selecting the best terms in searches.

2. Navigate to the EBI's OLS.

- Navigate to EBI's OLS in your browser of choice by clicking https://www.ebi.ac.uk/ols/index.

3. Search and identify standardized terms.

- Enter terms in the search bar and explore the results.
    - Try searching:
        - Province/state you are from
        - Your favourite food
        - A thing you can see from your window
        - "Cordyceps" (from The Last of Us)

- Record the best ontology term (label and ID e.g. pizza [FOODON:00003928]) for each item you searched in a text editor of your choice
- Consider the following when identifying ontology terms:
    - Does the term match sound like what I'm looking for?
    - Is the term being defined by an ontology that makes sense with my use case?
    - Does the definition sound right?
    - Is the term specific enough? Too specific?
    - Is the term reused in many different ontologies?

4. Participate in the group discussion about the activity.

- Reflect on your experience searching for term matches using EBI-OLS.
    - Was it easy to use?
    - Did you feel like the definitions were appropriate?
    - Was the hierarchy that was presented help you understand how terms were related to each other in the ontology?

Further Learning:

There are other ontology look-up services that you can explore:

OntoBee (includes all OBO Foundry ontologies)
BioPortal (>1000 indexed biomedical ontologies)

Learn more about how different ontologies are developed and used for infectious disease genomic epidemiology by reading these selected papers:

1. Genomic Epidemiology Ontology (GenEpiO): Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance
2. Food Ontology (FoodOn): FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration
3. Antimicrobial Resistance Ontology (ARO): CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database

Part 4 - Exploring the WHO's recommendations for pathogen genomics data sharing

Participants will learn about global efforts to encourage pathogen genomics data sharing for surveillance and pandemic preparedness. This exercise will explore the WHO's recommendations through a facilitated dialogue between the instructor and learners.

1. Listen to the instructor's overview of the WHO's 12 Guiding Principles.

- The WHO's 12 recommendations for pathogen genome data sharing include:
    1. Capacity building
    2. Collaboration and cooperation
    3. High quality, reproducible data
    4. Global and regional representativeness
    5. Timeliness
    6. Acknowledgement and intellectual credit
    7. Equitable access to health technologies as a benefit
    8. As open as possible and as closed as necessary
    9. Interoperability and relevance for national, regional and global decision-makers
    10. Trustworthiness and ease of use
    11. Transparency
    12. Compliance and enforcement

2. Participate in the group discussion.

After reviewing the WHO's guiding principles, join the discussion about what the principles mean for data generators and data users.
Consider the following:

1. What are the benefits of open data sharing vs controlled data sharing?
2. What are some of the risks of data sharing?

3. What types of data should be prioritized for sharing?
4. How can data standards facilitate data sharing?

Further learning:

Read the WHO's Guiding Principles for Pathogen Genome Sharing document in its entirety: https://www.who.int/publications/i/item/9789240061743

Read the WHO's 10 year strategic plan for global pathogen genomic surveillance: https://www.who.int/news/item/30-03-2022-who-releases-10-year-strategy-for-genomic-surveillance-of-pathogens

Resource Summary

1. Module 3 Lecture Slides: https://github.com/bioinformaticsdotca/IDE_2023/blob/main/module3/IDE_2023_DataCurationAndSharingLecture_Griffiths.pdf
2. DataHarmonizer (Pathogen Genomics Package): https://github.com/cidgoh/pathogen-genomics-package/releases/tag/PHPv2.0.3
3. Curation SOP: CBW Metadata Curation SOP_1.0
4. Public health genomic surveillance contextual data scenarios document: CBW2023 Contextual Data Curation Scenarios
5. Public repository submission protocols: https://www.protocols.io/workspaces/pha4ge
6. SARS-CoV-2 contextual data specification: Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package
7. EBI Ontology Look-up Service: https://www.ebi.ac.uk/ols/index
8. OntoBee Ontology Look-up Service
9. BioPortal Ontology Look-up Service
10. Genomic Epidemiology Ontology (GenEpiO): Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance
11. Food Ontology (FoodOn): FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration
12. Antimicrobial Resistance Ontology (ARO): CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database
13. WHO's Guiding Principles for Pathogen Genome Sharing document in its entirety: https://www.who.int/publications/i/item/9789240061743
14. WHO's 10 year strategic plan for global pathogen genomic surveillance: https://www.who.int/news/item/30-03-2022-who-releases-10-year-strategy-for-genomic-surveillance-of-pathogens