Infectious Disease Genomic Epidemiology - Data curation and sharing lab instructions

May 2024

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) An **exploratory exercise** to review the content and structure of contextual data within public repository records (NCBI).
2) A **data extraction exercise** to identify important data elements from mock public health data descriptions/scenarios.
3) A **demonstration** of a data curation and validation tool called the DataHarmonizer, followed by exercises to explore its curation features and functionality.
4) A **data curation exercise** where participants practice interpreting and structuring NCBI BioSample contextual data using the DataHarmonizer. Curated data will then be transformed into a GISAID repository submission format for comparison.
5) *Time permitting: 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 - Review existing data in NCBI BioSample.

Participants will familiarize themselves with the NCBI BioSample database, and compare the different types of contextual data stored within this repository.

Instructions:

1. Navigate to the NCBI BioSample database.

- Open your web browser and go to the NCBI BioSample website at https://www.ncbi.nlm.nih.gov/biosample/.

2. Search for the samples

- In this exercise we will be searching for the following samples:
    - hcov-19/usa/WI-94/2020
    - Sb3-tyagnc
    - SAMN14563388
- Locate the search bar at the top of the page and enter the first sample name in the search bar.
- The search results page will display any matching BioSample entries related to the sample identifier you entered. In this instance, there is only one exact match so the search function automatically opens the full BioSample record.

3. Review the contextual data and answer the following questions

- Look at the contextual data for each of these samples (such as sample name, organism, host, disease etc.), notice there is variety in the amount of data provided across the samples.
- Consider the following questions:
    - Are there any common elements or fields across the records?
    - Can you identify any potential inconsistencies or ambiguities in the non-standardized records?
    - What challenges might researchers face when trying to integrate or utilize this data?
    - How might standardization benefit both data producers and data consumers?
    - What steps could be taken to encourage greater adherence to data standards within the scientific community?
- If you have time, feel free to search for your own pathogens of interest.

Additional resources

- These examples all contain metadata related for NCBI BioSamples, in addition there are a number of other databases that ca store relevant contextual data, for example:

- **Genome Browser** https://www.ncbi.nlm.nih.gov/datasets/genome: Search and visualize annotated genomes for specific organisms or taxa. See results for Influenza A as an example.
- **Taxonomy Browser** https://www.ncbi.nlm.nih.gov/datasets/taxonomy/tree: Explore the hierarchical classification of organisms to understand their evolutionary relationships.
- **NCBI Virus** https://www.ncbi.nlm.nih.gov/labs/virus/vssi: Access virus genomic sequences and related data for virology research and study. Look at taxid:10239 as an example.
- **SRA Advanced Search** https://www.ncbi.nlm.nih.gov/sra/advanced: Perform detailed searches for sequencing data based on a variety of metadata and sequencing attributes.
- **Pathogen Detection** https://www.ncbi.nlm.nih.gov/pathogens: Investigate genomic data of bacterial and viral pathogens to track outbreaks and study antimicrobial resistance.

Part 2 - Data extraction from data descriptions

Participants will be provided with a set of scenarios of mock public health pandemic contextual data descriptions, and asked to parse out the important contextual data. All the materials can be found under the module 3 subdirectory on the github repository at https://github.com/bioinformaticsdotca/IDE_2024.

Instructions:

1. Review Scenarios:

- Scenarios doc: IDE 2024 Contextual Data Curation Scenarios
- Read through the provided mock scenarios carefully. Each scenario describes a set of contextual data related to the COVID-19 public health pandemic, with the first relating to a **clinical sample** while the second scenario describes a **wastewater sample**.

2. Identify Key Data Points:

- Identify all key data elements that are relevant to genomic epidemiology from each scenario (you can either pull this out into a spreadsheet, capture in a text editor or download the scenarios and highlight the key information).
- Focus on data such as pathogen characteristics, transmission data, demographic information, geographic locations, and any genomics data provided.
- Consider the following:
  - How would you structure this data, to make it as informative as possible?
  - What is the key information that needs to be captured to make data reusable and interoperable?

Part 3 - A demo of the DataHarmonizer.

The DataHarmonizer is a tool which facilitates the application of data standards. The instructor will provide a brief overview of the DataHarmonizer tool, with examples of how to enter, validate and transform data for repository submission. Participants are encouraged to download the DataHarmonizer and follow the instructor as we review how the mock public health scenarios presented in part 2 would be curated using this tool. All the materials (except for the DataHarmonizer tool) can be found under the module 3 subdirectory on the github repository at https://github.com/bioinformaticsdotca/IDE_2024.

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/PGPv5.1.0.
- Extract the zip file's contents, and save in an appropriate directory.
- Navigate into the extracted folder and open (double click) "**main.html"**. The validator application will open in your default browser[1].
- Note that in the address bar, the local address for the app is stored. The application is local to your machine and no data is shared online.
- The CanCOGeN[2] template will open as a default setting.

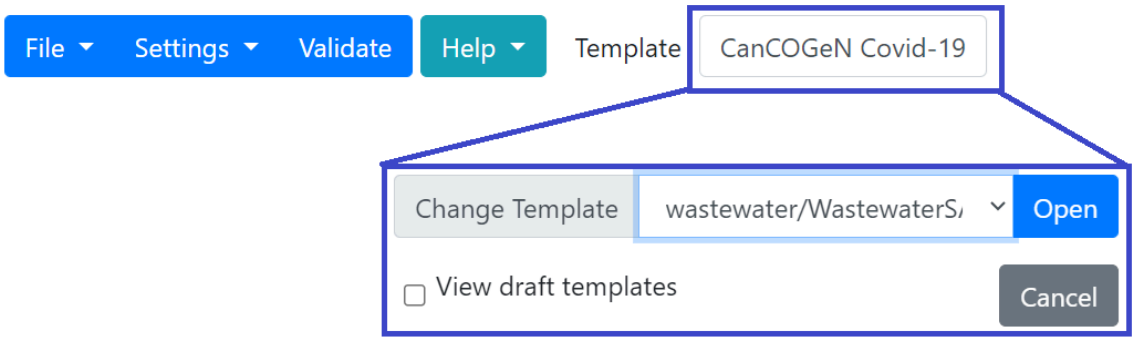
2. Review the public health contextual data scenarios and download the answer keys .

- Go back to the scenarios that you read in **Part 2**. Review the contextual data you identified and consider how this fits into the app based on the instruction provided in the demo.
- Download the excel files IDE 2024 clinical SC2 answer key and IDE 2024 wastewater SC2 answer key from the github repository (we will need these to open in the DataHarmonizer). You can find these under the module 3 subdirectory

3. Review the Wastewater SARS-CoV-2 scenario data in the DataHarmonizer app and validate the data.

- For this scenario we will need to use the template appropriate for wastewater samples. Navigate to the top of the screen and select the templates button and the **wastewaterSARS-CoV-2** template from the dropdown.

[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



- Go to **File -> Open** and select the downloaded IDE 2024 wastewater SC2 answer key.
- Follow along with the demonstration to learn more about the different fields, including an extra module for environmental conditions and measurements.
- To view this module independently, go to **Settings -> Modules -> Environmental conditions and measurements.**
- Pick an environmental measurement field that is currently empty. Double click on the field header to learn more about its definition and how to structure information in that field. Add some data using the guidance.
- When complete, click "**Validate**" in the menu to check for errors and missing information.
- Make any corrections necessary.

4. Review the Clinical SARS-CoV-2 scenario in the DataHarmonizer app and validate the data.

- To do this, we first need to use a different template. Navigate to the template icon and change template to **CanCOGen**.
- Go to **File -> Open** and select the downloaded IDE 2024 clinical SC2 answer key.
- Follow along with the demonstration to learn more about the different fields.
- To review required and recommended fields only go to **Settings -> Required and Recommended columns.**
- Imagine you've been given the sequencing data by the data provider, add this information to the template.
- **Validate** as before and make any corrections necessary.

Part 4 - Curating contextual data using the DataHarmonizer.

Participants will practice using the DataHarmonizer application. Participants will have the opportunity to curate NCBI records they found in part 1. 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. Curate the NCBI examples from Part 1

- Review the DataHarmonizer Curation SOP: IDE Metadata Curation SOP_1.0
- Consider the NCBI entries we searched for in Part 1, and the inconsistencies in how the data was structured. See the metadata variability examples in github for a summary of the records.
- Curate these records using the DataHarmonizer tool. To do this we recommend using the **PHA4GE** template.
- Enter as much of the information into the template as possible. Tip: Change to view to 'required and recommended' to limit the columns for ease of use, or enter information module by module.
- Consult the curation SOP for examples and additional guidance.
- Validate and check for errors.

2. 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?

3. Export the data in GISAID[3] format.

- Interoperability is an important factor in data sharing. Using a tool such as the DH we can export for different standards, which facilitates data re-use. As an example, let's **export** the data you've just curated in a GISAID ready format.
- To export go to **File -> Export To -> Format**, select the **GISAID** template and name the file for download.
- There are a number of different export options which generate submission ready formats for well known public repositories. If you have time, try exporting in these formats and explore the differences.

[3] Global Initiative on Sharing Avian Influenza Data (GISAID) - www.gisaid.org

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 5* - Exploring the WHO's recommendations for pathogen genomics data sharing

***Time permitting**

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

[4] National Centre for Biotechnology Information (NCBI) - https://www.ncbi.nlm.nih.gov/
[5] European Nucleotide Archive (ENA) - www.ebi.ac.uk/ena

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?

Additional resources:

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

Additional learning resources - using ontology look-up services to standardize data

You will have noticed that many of the DataHarmonizer templates contain controlled vocabulary for different fields. These terms are based on ontologies. Ontology search tools such as European Bioinformatics Institute's ontology look-up service (EBI-OLS) can be used 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.

Instructions:

1. 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.

2. Search and identify standardized terms.

- Enter any term 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?

3. Learning considerations.

- 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

Resource Summary

1. Module 3 Lecture Slides: https://github.com/bioinformaticsdotca/IDE_2024/blob/main/module3/IDE_2024_DataCurationAndSharingLecture_Griffiths.pdf
2. DataHarmonizer (Pathogen Genomics Package): https://github.com/cidgoh/pathogen-genomics-package/releases/tag/PGPv5.1.0.

3. Curation SOP: CBW Metadata Curation SOP_1.0
4. Public health genomic surveillance contextual data scenarios document: https://github.com/bioinformaticsdotca/IDE_2024/blob/main/module3/IDE_2024_Contextual_Data_Curation_Scenarios.pdf
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