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

PHA4GECon 2025: Data Needs Assessment

Identify the data needs and get used to asking questions!

Project Name: <> Data Standards Development
Interviewee Name: Name
Affiliation / Role:
Date & Time: April 2025
Note Taker:
Consent:

☐ Video ☐ Audio ☐ Video & Audio ☐ Notes only

Overview

Summary of the specification project aims and context

The <team name> has been engaged to develop a data specification for the <project name> project. The goal of this project is to <project goals>. This requires identifying the contextual data lifecycle including, what data is being collected and how, how it is managed and used, and any downstream processing or sharing.

Purpose of the interview

To identify current data practises, data size, type and variety, as well as the processes involved to inform the needs assessment component of developing a data standard. We aim to enhance collaboration, improve data quality, and support informed decision-making in <project>

Stakeholder context

Instructions: Provide a brief description of the interviewee's background and role in the data lifecycle. Include their organisation type (e.g., research, government, policy, clinical, agriculture) and what part(s) of the data pipeline they contribute to (e.g., sample collection, analysis, data use, dissemination). Describe any known partners or networks they collaborate with, and the types of data they handle.

Job Title	
Organisation type	(e.g. Research, Government, Policy, Agriculture etc)
Role in data lifecycle	(e.g. sample/data collection, curation, analysis, use, dissemination etc)
Types of data currently collected	

Interview

Scope and audience

Instructions: Ask the interviewee to describe the aims of their work or project.

1. What are the questions your project/initiative is trying to answer (e.g. surveillance, research, monitoring a specific pathogen or target)?

2. Who are your partners?

Sampling Strategy

3. What kinds of samples are collected (host, environmental, food, wastewater, etc.) and what is the sampling frequency and duration?

4. Are you using any previously collected samples?

5. What contextual data is being collected alongside the samples (e.g. hosts demographics, physico-chemical measurements, weather, location, proximal location, presampling activities, collection method, collection device etc)

6. Do you have a sample plan? Is the sample plan, experimental design, or any other material accessible?

Specimen processing

7. Are the samples subjected to any special processing before sequencing (e.g. filtering, enrichment, culturing, etc)?

8. If samples are cultured, what kinds of media and conditions are being used?

Sample receipt and storage

9. How are samples stored and for how long (e.g. temperature, duration, location tracking)?

10. Are there receipt processes (e.g., documentation, quality checks, chain-of-custody)?

Library preparation

11. What are the primary sequencing assays (e.g WGS, WRS, amplicon etc)?

12. What library preparation or enrichment methods are applied?

13. Are you using commercial library kits or in-house developed methods?

14. What are your controls?

Sequencing and Bioinformatics

15. What sequencing instruments are being used?

16. What bioinformatics tools, pipelines and reference databases are being used? Are there any community workflows or tailored scripts?

Pipelines	
Tools	
Databases	

17. How is quality control done and recorded?

Data Management

18. How is the contextual data being captured (e.g. spreadsheets, LIMS, RedCap etc)?

19. Are any data standards being used to structure the data? If not, would there be an interest in using a standard?

20. Where is the data being stored? Locally, public repositories, network databases?

21. What are the data sharing workflows - where does it need to go, how does it get there? Are there any restrictions?

22. During data flow, will the data need to be transformed?

23. Are there any governance, policy or privacy challenges?

24. Is there a plan to share the data publicly? If not, why?

Associated Data Types

25. Are there any other data types or characterizations available for the samples or sequences that were not previously mentioned as contextual data?

26. Which of these would be useful to indicate availability alongside the sequencing data?

Future Work (anticipating future needs)

27. What data types would you want to collect/foresee integrating in the future?

28. What are your main contextual data challenges?

29. What would be on your contextual data wishlist?

Summary

Data ecosystem	Details
Types of data currently collected	
Current data workflow	
Systems, standards and/or tools currently used	
Key outputs/reports that data supports	

Challenges	Details
Pain points	
Missing data fields	
Integration challenges	
Data governance and permissions	