Experiments Metadata Checklist: PHA4GE DSWG Feedback

Reference materials

- Overview (webpage)
- Experiments metadata checklist
- GitHub repository

This feedback was gathered from members of the Public Health Alliance for Genomic Epidemiology (PHA4GE) Data Structures Working Group (DSWG), who have been following the development of the GA4GH Experiments Metadata Checklist and reviewing its applicability in public health and pathogen genomics contexts. The comments reflect both **technical considerations**, ensuring the data model is unambiguous, comprehensive, and interoperable and **implementation guidance** to support smooth adoption by groups and repositories. Feedback draws on PHA4GE's experience in developing standards that must work across diverse settings, including regulated environments, large-scale research initiatives, and smaller public health laboratories.

Technical Feedback

Relating to the underlying model, field definitions, and data quality enforcement.

- Identifier uniqueness & examples
    - If identifiers in `identifiers.md` must be truly unique, this requirement should be explicitly stated.
    - Examples should reflect actual accession formats if "accession number" is specified (e.g., `SAMN)`
    - Consider separating:
        - `sample_id` (user-provided, internal)
        - `sample_accession_id` (repository-assigned, e.g., INSDC) to maintain provenance. In PHA4GE data standards we use `insdc_sample_accession.`
- Field redundancy & scope
    - Potential overlap among `molecule_type`, `assay_type` and `experiment_type`. Can experiment_type be derived from assay_type? Clarify necessity and derivability to avoid duplication (it seems the values in assay_type are just subclasses of experiment_type). PHA4GE captures sequencing_assay_type only.
- Coverage of new read types
    - `library_layout` is limited to paired-end (PE) and single-end (SE) → expand to include newer long-read and emerging sequencing technologies?

- Addition of library prep kit information/library prep protocol URL, contact info for lab performing sequencing
    - Many fields are devoted to the type of assay/molecule but there are few fields that capture the details of the assay itself
    - Unclear if the sequencing protocol includes library prep
- Picklists & ontology compliance
    - Make clear that examples are *illustrative*, not exhaustive picklists.
    - Define a validation approach for CURIE formats (e.g., `GENEPIO_0001234` vs `GENEPIO:0001234`).
    - Specify authoritative ontologies *and the relevant branches* for each field.
- Beyond Experimental/SRA requirements
    - For pathogen genomics, consensus sequences/assemblies are very often the type of sequence data that is shared/used. Excluding the capture of these methods from the standard may limit its utility for public health pathogen genomics use.
- Mapping/alignment with PHA4GE sequencing standards
    - PHA4GE has developed additional fields for capture of sequencing methods. Would be beneficial for the community to see efforts to reuse/create interoperability between host/pathogen standards.

Implementation & Adoption Feedback

Supporting correct and consistent application of the standard.

1. Clearer demonstration of innovation/novelty
    - Highlight the new fields/terms (currently difficult to intuit, looks a lot like pre-existing INSDC requirements)
2. Clearer documentation of identifier use
    - Show mappings between internal IDs and repository accessions, with examples of both captured and linked.
    - Alongside a precise definition, provide guidance showing how to capture both internal and repository-assigned IDs, with real-world examples from different repository systems.
3. Provide additional supportive guidance as well as definitions
    - Add a guidance column explaining how to apply the definition in practice. Can include edge cases, common pitfalls and/or recommended values or sources. Useful for implementers who are mapping from existing metadata or building submission tools.
4. Picklist vs free text
    - Explicitly state that example values are not complete lists of allowable options.
    - Provide guidance or tooling for generating valid CURIEs.
5. Future: Schema & validation support
    - Offer JSON Schema or equivalent validation tooling early to help implementers.
    - Include ready-to-use validation scripts/checkers in the GitHub repository.