Data Needs Assessment Questionnaire

Date

Scope

- What are the primary questions your project/initiative is trying to answer?
  Research vs surveillance? Pathogen/target? Assay?

- Who are your collaborators/partners (informs data flow, sharing, harmonization needs)

Sampling Strategy

- What kinds of samples are you collecting?
  Hosts (anatomical materials/parts/body products)/ environments (materials/sites)

- What are the collection devices and methods being used?
  Follow-up if required: Are these standard tools in this area of research or adapted for this specific project?

- Are you using any previously collected samples?
  Speaks to purpose of sampling/sample bias, as well as data reuse

- What contextual data associated with the samples is being collected/considered?
  - Hosts - age/gender/exposures
  - Physico-chemical attributes? Weather?
  - Land use, water table and/or tidal data, zone info (benthic/pelagic)?
  - Presampling activities
  - Experimental interventions?
  - Range of geo-locs and other geo-spatial data
  - Sampling frequency and duration

- Is there a documented sample plan for this project? Is the sample plan, experimental design, or any other material available online or shareable?

Specimen processing and Storage

- Are the samples subjected to any special processing before sequencing libraries are prepared?
  Filtering, enrichment, culturing, etc

- For cultured samples what kinds of media and conditions are being used?

- How are samples stored, and for how long (cold chain, comparability)?

Library Preparation

- What are your methods for extraction, and any types of enrichment?

- How are sequencing libraries being generated?

- What quality/controls elements are included?
  Endogenous controls, replicates, synthetic constructs/data etc.

Sequencing and Bioinformatics

- What sequencing instruments are being used?

- What bioinformatics processing is being performed

    - Are workflows custom scripts or established/community-developed?

    - Specifically, what (if any) is being used for:
        - Raw sequence processing and/or dehosting

        - Assembly or mapping

        - Variant calling or lineage assignment

- QC assessment
- Databases

Associated Data Types

- Are there other data types or characterizations available for the samples/sequences?
  - Physico-chemical properties*:*
    Temp, pH, oxygen, salinity, nitrogen, phosphorous etc
  - Biological properties:
    Enzymatic, staining, marker identification, colony counts, AMR, typing (serotype), matter composition, mollusc shell length, therapeutic history, etc

Data Management

- How is the contextual data being captured?
  Spreadsheets, LIMS, RedCap, other
- Are any data standards being used to structure the data?
- If no, would your partners be interested in using international standards?
- Where is the data being stored?
  Locally, public repositories, network databases?
- What is the data flow?
  From collection to storage. Where does it need to go, how does it get there
- During data flow, will the data need to be transformed?

Data Sharing & Governance

- Among your network(s), are there any data use limitations/restrictions
  Consider beneficial data use licensing/tagging
    - Due to limitations/biases of the data?

    - Due to Data governance?

- Is there a plan to share the data publicly? If no, why? What are the perceived challenges?

Future Work (anticipating future needs)

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

- What are your main contextual data challenges?

- What would be on your contextual data wishlist (if resources were unlimited)?