GRDI Contextual Data Guidance - Best Practices for Sample and Isolate Identifiers

Aug 2021

Sample Identifiers for Data Sharing

A sample is considered to be the material originally sampled outside the laboratory. Identifiers generated for internal sample tracking can conform to lab conventions with little friction, however, when data is shared with trusted partners or public repositories, data from different sources must be integrated. In these contexts, different sample identifier conventions can quickly become problematic.

Best practices for sample/isolate identifiers include containing sufficient complexity to avoid identifier clash (uniqueness), tracking of original and “alternative” (renamed) identifiers for traceability and establishing chain of custody, avoiding the inclusion of information that could violate institutional privacy policies (no metadata e.g. dates, geographical locations, source attributes identifiers) or non-alphanumeric symbols (with the exception of underscores).

Ideally, identifiers are issued from a central authority for tracking and indexing purposes, and to ensure that every identifier is truly unique. In the absence of a central authority for issuing identifiers across Canadian agencies, a format for identifiers is recommended below.

Recommendations:

1. Identifiers should be created in this format:
   “4 digit agency identifier”_”short local lab identifier”_alphanumeric unique sample ID”
   e.g. CFIA_CC_ABX123
2. Original sample identifiers should always be included in contextual data records to establish chain of custody and track provenance. Original sample identifiers should be recorded as the “specimen_collector_sample_ID”.
3. Samples shared between agencies that are assigned subsequent identifiers should be tracked as the “alternative sample ID”.
4. Sample identifiers should not contain any contextual data e.g. dates, geographical locations, source attributes, nor should they contain symbols, dashes, or the capital letter “O” (Exception: if “O” is in a standardized agency identifier e.g. DFOC).
5. Laboratories should have a protocol for generating and tracking identifiers. The mechanism for generating and tracking identifiers is at the discretion of the lab.
6. Agency and local lab identifiers should be used consistently across agencies and labs. Identifiers for **agencies** are suggested below.

Agency	Suggested Agency Identifier
Canadian Food Inspection Agency	CFIA
Public Health Agency of Canada	PHAC
Agriculture and Agri-Food Canada	AAFC

Environment Canada	ECCC
Health Canada	HCAN
Fisheries and Oceans Canada	DFOC
National Research Council Canada	NRCC

7. Laboratories should reach consensus on what their lab’s identifier should be, and then use it consistently when sharing data.
8. Isolates derived from samples should be ascribed their own identifiers, called the “isolate_ID”.
9. Laboratory or type culture collection reference strains should be ascribed their own identifiers, called the “strain_ID”.
10. When submitting to the INSDC (i.e., NCBI), provide the “specimen_collector_sample_ID” as the “sample_name”. Provide the isolate_ID as the “isolate” (unless a reference strain was sequenced, in which case provide the “strain_ID” as the “isolate”). A BioSample should always be created for every sample/isolate sequenced. Multiple sequence datasets can be submitted and linked to a BioSample.