Biospecimen Metadata Requirements

Author: *Kristen Anton*

Updated: *July 17, 2024*

Status: Approved

Terminology:

This document uses the following terms from IEFT RFC 2119

- **MUST / REQUIRED / SHALL**: ✅ (denotes absolute requirement)
- **MUST NOT / SHALL NOT**: ❌ (denotes absolute prohibition)
- **SHOULD / RECOMMENDED**: 🌟 (denotes recommendation)
- **SHOULD NOT / NOT RECOMMENDED**: 🤦‍♂️ (denotes not recommended)
- **MAY / OPTIONAL**: 🤷‍♀️ (denotes optional)

Biospecimen Metadata

Biospecimen Metadata is designed to capture a description of each biospecimen contributed by an HTAN participant from two perspectives: physical (for instance, sample capture and processing details, physical dimensions, and shipping information); and histological (for example, cell type composition of the specimen and histologic morphology).

BIospecimen Metadata is captured in a single Tier, therefore requiring one file to capture all biospecimen annotations.

Biospecimen data is captured both for parent biospecimens, or the primary specimens collected directly from the HTAN participant (e.g., tissue, blood, sputum), and for derivative specimens (e.g., DNA, RNA, tissue section, tissue block). For each derivative biospecimen, the HTAN Parent ID is required, thus maintaining the chain of derivation.

Biospecimen identifiers are linked to the HTAN participant and to all molecular profiling data files generated from the biospecimen. This allows critical linkage of data and/or metadata across the repository.

The biospecimen metadata manifest is accessible through the HTAN Data Curator App and is downloadable in Google sheet format. The manifest includes a mix of required (blue cell) and optional data elements (white cell). Conditional requirements also exist; as data elements become required the

cells become blue shaded in the manifest. These requirements are also denoted in the HTAN data standards documentation.

Standardized controlled vocabulary exists for many biospecimen data elements. The controlled vocabulary facilitates data harmonization and search across the HTAN dataset. Standards must be followed for the manifest to pass the validation check and be accepted into the HTAN dataset.

All HTAN data elements that capture time points are expressed as “days to,” to obfuscate dates while allowing for longitudinal timelines and relative data analysis. **The index date for all HTAN dates is the participant date of birth.** For example, a specimen collected at X=365 days is collected 365 days after the participant’s date of birth.

The data standards allow for multiple responses for many data elements. This information may be found in the “Validation” column of the HTAN Data Model documentation. Multiple responses are submitted delimited by commas.

✅ One biospecimen record **is REQUIRED** for each biospecimen.

✅ All dates **MUST** be converted to days from index, where the index is the participant date of birth.