# File Manifest Specification

A file manifest listing files representing data about generic scientific samples.

1. The manifest will follow a tabular logical data model, with one table. If a name for the table is needed, file\_manifest or **File manifest** may be used.
2. Each record in the table should pertain to exactly 1 file, meaning an exact byte string materially realized in electronic storage in some standard file system (e.g. ext4, FAT32, NTFS).
3. Each record should include the following 11 fields (if marked "Optional", the value may be empty to indicate that the information is not known or not applicable; otherwise it must not be empty):

|  |  |  |  |
| --- | --- | --- | --- |
| Name of field | Machine-friendly name |  | Description of intended string contents |
| **File ID** | file\_id | *Required* | This should be a primary identifier for the file in use by the creator of the manifest (hereafter called the "curator"). |
| **Project ID** | project\_id | *Optional* | This identifier is preferably an identifier, suitable for circulation, that identifies a logical unit similar to a “project”, “dataset”, or “data collection”, whatever such unit is most readily available from the point of view of the curator. Ideally this string is textual and somewhat human-readable, as opposed to numeric or hash-like. This field is optional to allow for the usage of this field as a convenience of data organization, in case it is helpful for the curator. A full account of asset-to-“project” relationships (and other “project” information) is considered out of scope for this specification. |
| **File name** | file\_name | *Optional* | If the record pertains to a file, this should ideally be the file’s basename (i.e. without path information, but with file extension if this extension is part of the filename). This field is optional to allow for the case that the curator deems the values to be unsuitable for circulation, either because they are subject to change or for any other reason. |
| **Sample ID** | sample\_id | *Optional* | This identifier should identify a specimen that the record (digital asset) is about, at whatever granularity the data submitter cares to provide, within the namespace defined by the “Project ID” value (which must therefore be provided, whenever “Sample ID” is). Note that specification of this granularity (e.g. cell, brain, organism) is the purview of a sample or specimen manifest, not this manifest. Usage of this field ideally implies that the “Project ID” provides enough information to identify a group of people (e.g. a lab) some subset of which were responsible for the generation of these “Sample ID”s. |
| **Availability** | availability | *Optional* | An indication of the degree of public availability the data generator / curator intend for the file asset. Commonly, this value will be “Public”, but special cases may be indicated by alternatives like “Private”, “Controlled access by … ”. |
| **URL** | url | *Optional* | A Uniform Resource Locator (URL), resolvable within some subnetwork, that retrieves a document which is an exact representation of the digital asset. |
| **Network** | network | *Optional* | A description of the network or subnetwork within which the direct URL may be resolved. E.g. "Public internet" or "Internal corporate HPC". |
| **Data type** | data\_type | *Required* | A relatively free-form string describing the data type of the asset. For example, “SAM file -- Sequence Alignment Map”. More detail should be included commensurate with feasibility and disambiguation utility. |
| **Checksum** | checksum | *Required* | The value of this field should be the hash of the file byte string with respect to the checksum scheme named in the “Checksum scheme” field. Use the ASCII representation of this checksum which is standard for that scheme (e.g. for SHA256, a 64-character lowercase alphanumeric string). |
| **Checksum scheme** | checksum\_scheme | *Required* | Indication of the cryptographic hashing scheme used to verify the integrity of the file. Most likely “SHA256” or “MD5”. Alternatives may be used according to need. |
| **Size** | size | *Required* | The value of this field should be the integer number of bytes in this file byte string. |

1. In all non-empty fields, the only characters allowed are the printable ASCII characters. Note that this excludes hyphen-like characters beyond the ASCII hyphen “-”, as well as so-called smart quotation symbols. No trailing or prepended spaces are allowed. Each field must contain 2 or more characters. A regex pattern enforcing these requirements is:  
    ^[!-~][ -~]\*[!-~]$