# BICCN File Manifest Specification for Data Archive Communication

# Version History

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Description of Change | Author |
| 0.1 | 01/13/2021 | Initial draft, to be circulated to File Metadata Task Force | Jimmy Mathews |
| 0.2 | 01/14/2021 | Added comments during discussion | File Metadata Task Force |
| 0.3 | 01/15/2021 | Edited to include suggestions made by task force members | Jimmy Mathews |
| 0.4 | 01/15/2021 | Comments – to occur | BICCN Infrastructure Working Group |
| 1.0 | 03/16/2021 | Prepared for initial release | Jimmy Mathews |

## File Manifest Specification Notes

1. Each R24 archive will produce its own asset manifest. Provisionally this is planned to take place quarterly.
2. 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.
3. The manifest itself should be hosted somehow by the R24 archives. CSV/TSV file download is acceptable in the short-term. A web API is acceptable, if it simulates the flat table structure effectively.
4. Each record in the table should pertain to exactly 1 file, and that file must be uniquely determined by the table record.[[1]](#footnote-1)
5. Each record should include the 12 fields listed in the Table 1 below. (If marked "Optional", the value for that field may be empty to indicate that the information is not known or not applicable; otherwise it must not be empty).
6. 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. A regex pattern enforcing these requirements is: ^[!-~][ -~]\*[!-~]$ [[2]](#footnote-2)
7. The archives and data submitters jointly decide which file assets to record in the manifest. For example, they may adopt a policy that excludes certain deposited files from appearing on this "BICCN-scoped" manifest, for simplicity or convenience.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Name of field | Machine-friendly name |  | Description of intended string contents |
| **Asset ID** | asset\_id | *Required* | This digital asset identifier is preferably an archive-specific, archive-issued identifier, suitable for circulation. |
| **Project ID** | project\_id | *Optional (recommended if* ***Sample ID*** *present)* | This identifier is preferably an archive-specific, archive-issued 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 archive. 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 maintainer. A full account of asset-to-“project” relationships (and other “project” information) is considered out of scope for this specification. |
| **Asset name** | asset\_name | *Optional* | 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 archive and/or data submitter deems the values to be unsuitable for circulation, either because they are subject to change, or because no canonical name exists in the filesystem of storage, or for any other reason. |
| **Sample ID** | sample\_id | *Optional* | This identifier should identify a biological specimen that the file record 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 provided). 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’s members), some subset of which were responsible for the generation of these **Sample ID** values. |
| **Public availability** | public\_availability | *Optional* | An indication of the degree of public availability the data submitter and archive intend for the asset. Most commonly, this value will be “Public”, but archive- and project-specific special cases may be indicated by alternatives like “Private”, “Controlled access by … ”. |
| **URI** | uri | *Optional* | An identifier uniquely identifying the record (digital asset) using the syntactic conventions of a Uniform Resource Identifier (URI). In particular, the identifier should most likely begin with <https://dandiarchive.org/>, <https://nemoarchive.org/>, or <https://www.brainimagelibrary.org/> or similar. General URI and URL specifications can be found in [RFC 3986](https://tools.ietf.org/html/rfc3986). |
| **URL** | url | *Required if “URL direct” is not provided* | An internet-resolvable Uniform Resource Locator (URL) that retrieves a document which provides precise instructions for the user to obtain the asset (or otherwise redirects the client to such a document; see HTTP status codes 303 and 307). |
| **URL direct** | url\_direct | *Required if “URL” is not provided* | An internet-resolvable Uniform Resource Locator (URL) that retrieves a document which is an exact representation of the digital asset. Although this field theoretically enables programmatic access to raw assets, extensive such use is not recommended. Instead, a future system and specification should be used which is specifically designed to support such programmatic access (especially to satisfy bulk requests). |
| **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. For example, there are non-interoperable variants of SWC neuron reconstruction files which ought to be disambiguated here if it is feasible to do so. |
| **Checksum** | checksum | *Required* | The value of this field should be the hash of the byte string representing the contents of the file, 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 asset. Most likely “SHA256” or “MD5”. Alternatives may be used according to need. |
| **Size** | size | *Required* | The value of this field should be the decimal-formatted integer number of bytes in the byte string representation of the file. |

1. Prior versions of this document proposed an extension to directory contents, but that turned out to be too difficult to specify in exchange for too little benefit. [↑](#footnote-ref-1)
2. It was noted that this limitation is almost certainly too strict. Lifting this limitation, however, requires some consideration to determine the appropriate replacement. Surely not all bytestrings should be allowed. [↑](#footnote-ref-2)