# Metadata Extraction (CPP-009)

| | |
|---|---|
| **CPP-Identifier** | CPP-009 |
| **CPP-Label** | Metadata Extraction |
| **Author** | Bertrand Caron |
| **Contributors** | Juha Lehtonen |
| **Evaluators** | Matthew Addis, Maria Benauer, Fen Zhang |
| **Date of edition completed** | 29.08.2025 |
| **Change history** | **Comments** |
| Version 1.0 - 29.08.2025 | Milestone version |

# 1. Description of the CPP

The TDA extracts characteristics (such as size, image dimensions, video codec, audio run time, creating application).

## Inputs and outputs

| | | |
|---|---|---|
| Input(s) | | |
| Data | *File* or, in some specific cases[1], *Representation* | |
| Metadata | *Technical Metadata* (Format identifier) | |
| Documentation / guidance | Policies and detection methods for significant properties, risk properties and quality properties | |
| | Metadata recording Policy | |
| Output(s) | | |
| Metadata | *Technical Metadata* | |
| | *Provenance Metadata* | |
| | Optional | *Descriptive Metadata* |
| | | *Structural Metadata* |
| | | *Rights Metadata* |
| | | Errors and Warnings |

## Definition and scope

Metadata extraction is the process of analysing a *File* or a set of *Files* (i.e. a PREMIS Representation) by means of metadata extractor tools in order to retrieve its characteristics in an automated way. In the digital preservation community, this operation is sometimes referred to as "Characterisation". Within EDEN, however, characterisation is used to identify all operations aiming to extract properties from digital *Files* (CPP-008 **File Format Identification**, CPP-009 **Metadata Extraction**, and CPP-010 **File Format Validation**).

File properties as gathered through Metadata Extraction are generally considered "*Technical metadata*". However, by parsing the *File*, the process also extracts a wide range of its internal *Descriptive*, *Provenance* and *Rights metadata*.

Knowledge of these characteristics is a key requirement for many subsequent operations. In particular, it is essential for CPPs producing new *Representations* (i.e. CPP-026 **File Normalisation**, CPP-027 **File Repair**, CPP-014 **Format** Migration, and CPP-028 **Creation of Derivatives**) as they require further *Metadata* beyond the file format information as

[1] Metadata Extraction is generally performed at a *File* level; in some cases it has to be applied to a complex file structure that is not wrapped in a container *File*. For example, moving images stored as a sequence of DPX *Files* are handled as a whole by the metadata extractor tool MediaInfo.

provided by CPP-008. Audiovisual *Files* are the most obvious example: Most identification tools provide information about the only container format, while any of the operations mentioned above will need at least the video and audio stream format. This is equally true for all other data types: TIFF may contain image streams compressed by different algorithms, PDF 1.7 might be portfolios containing arbitrary *Files*, etc.

There are different types of extraction tools available: a) Generalist metadata extractor tools which are able to perform metadata extraction on a great variety of file formats of different content types (e.g. Exiftool), b) content-specific tools which cover most of the file formats for a specific content type (e.g. MediaInfo for AV *Files*), c) format-specific tools which are specialised on a particular format (e.g. EPUBcheck for EPUBs, metaflac for FLAC *Files*, etc.). Metadata extraction therefore relies on a format identification and is performed differently depending on the file format. TDAs might decide to apply several metadata extractor tools if a single tool cannot extract all required properties.

Metadata extraction - like any parsing operation - can fail and result in diagnostic error messages. These errors require systematic analysis to inform troubleshooting. In particular, the following issues may be detected by *Metadata* extraction errors:

- Encrypted *Files*;
- Truncated or broken *Files*;
- *Files*assigned incorrect file format information.

The output of the metadata extractor tools should be recorded in the *Information Package*, as *Technical*, *Descriptive*, *Rights* or *Provenance Metadata*. It may be recorded directly as-is, or mapped to a metadata standard according to the TDA policy on metadata recording.

# Process description

## Trigger event(s)

| Trigger event | CPP-identifier |
|---|---|
| Ingest | CPP-029 (Ingest) |
| Re-run of metadata extraction because of the release of a new metadata extractor tool or tool version | / |
| Verify the output of processes creating new *Files* or *Representations* | CPP-014 (File Migration), CPP-026 (File Normalisation), CPP-027 (File Repair), CPP-028 (Creation of Derivatives) |

## Step-by-step description

| No | Supplier | Input | Steps | Output | Customer |
|---|---|---|---|---|---|
| 1 | CPP-008 (File Format identification) | Format identifier | Select a suitable metadata extractor tool for the *File*(s), depending on its format identifier and on requirements from Significant Properties Definition, Data Quality Assessment and Risk Definition and Extraction | Metadata extractor tool | |
| | CPP-019 (Data Quality Assessment) | Significant, risk and quality properties policy and detection methods | | | |
| | CPP-022 (Significant Properties Definition) | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | If relevant: configure the tool settings - syntax (XML, JSON, CSV, etc.), format (e.g., EBUCore, PBCore, in case of an AV *File*) and verbosity level. | |
| | CPP-023 (Risk Definition and Extraction) | | | | |
| 2 | | *File* | Applying one or sometimes several metadata extractor tool(s) | Tool(s) output | |
| | | Metadata extractor tool configured according to selected settings | | Optional: errors and warnings | |
| 3 | | Errors and warnings | Analyse the errors and troubleshoot, e.g., by removing encryption | | |
| 4 | | Tool(s) output | Optional: Mapping of the extractor tool(s) output to metadata standard(s)[2] | Tool output in standard format(s) | |
| | | Policy on metadata recording | | | |
| 5 | | Tool output in a standard format | Recording the results in the *Information Package* (i.e. in | *Technical metadata*, optionally other types of | CPP-029 (Ingest) |

[2] The output of the extractor tool may be recorded as-is in the Information package, or may be mapped into a technical metadata standard (e.g., MIX for still images).

| | | | | | |
|---|---|---|---|---|---|
| | | | practice, in the digital archive database and/or in the physical *Information Package*) | metadata, recorded in the *Information Package* | optional: CPP-021 (AIP Versioning) |
| 6 | | | Document event and its datetime | *Provenance metadata* | |

### Rationale(s)[3] and worst case(s)

| Rationale | Impact of inaction or failure of the process |
|---|---|
| All preservation actions beyond bit-level preservation are based on a comprehensive understanding of the *File's* characteristics. | *Files* of poor quality may be unidentified. |
| | Derivatives may be unadapted to the end users' needs. |
| | The result of a migration may be partial, as some parts of the source *File* may not have been identified, thus not been copied to the target *File*. |
| Metadata extraction involves accessing the contents of the *File*. Hence it is an essential means to detect problematic *Files* (including errors like: corrupted, non-conformant to format specification, encrypted or password protected, wrong file format identification etc.). | Problematic *Files* are not detected. |

## 2. Dependencies and relationships with other CPPs

### Dependencies

| CPP-ID | CPP-Title | Relationship description |
|---|---|---|
| CPP-008 | File Format identification | The selection of an appropriate extractor tool depends on file format information. |
| CPP-019 | Data Quality Assessment | The selection of an appropriate extractor tool depends on requirements from the Data Quality Assessment CPP. |
| CPP-022 | Significant Properties Definition | The selection of an appropriate extractor tool depends on requirements from the Significant Properties Definition CPP. |
| CPP-023 | Risk Definition and Extraction | The selection of an appropriate extractor tool depends on requirements from the Risk Definition and Extraction CPP. |

[3] Term derived from PREMIS.

## Other relations

| Relation | CPP-ID | CPP-Title | Relationship description |
|---|---|---|---|
| Required by | CPP-012 | Risk Mitigation | Preservation actions (i.e. migration, emulation) in the storage depend on the identification of *Files* that share the same properties. |
| Required by | CPP-014 | File Migration | File format identification is generally limited to an indication of the container format, while migration can apply to any property of the *Files*. Technical metadata extraction is required to both assess the compliance of files format to the Archive's format policy and control the outcome of the migration. |
| Required by | CPP-016 | Metadata Ingest and Management | Any metadata that was extracted from the *File* needs to be stored, searchable and retrievable |
| Required by | CPP-019 | Data Quality Assessment | Metadata extraction returns *Metadata* that are used to assess the *File* quality (e.g. for an audiovisual *File* quality assessment may rely on *Metadata* such as bit depth, sampling frequency, etc.) |
| Required by | CPP-023 | Risk Definition and Extraction | Metadata extraction returns *Metadata* that are used to identify preservation threats (e.g. for a PDF, the presence of an open password). |
| Required by | CPP-024 | Enabling Discovery | Some *Metadata* provided to the consumer must have been extracted from the *Files*. |
| Required by | CPP-029 | Ingest | Metadata extraction is one of the core processes that must be performed as part of Ingest. |
| May be required by | CPP-010 | File Format Validation | Depending on the precision of the format registry used in the format identification process, the resulting information may be insufficient for selecting the right validation tool.<br>In such cases, additional *Metadata* from an extraction tool may be required. For example, if an organisation uses Unix *File* as its identification tool, which does not |

| | | | |
|---|---|---|---|
| | | | distinguish between different PDF “flavours”, and wants to validate PDF/A against the PDF/A standard.<br>In that case, metadata extraction will be necessary to identify the conformance level and select veraPDF as the suitable validation tool. |
| May be required by | CPP-021 | AIP Versioning | The documented event, datetime, and *Provenance metadata* from the metadata extraction may be required by AIP Versioning. |
| May be required by | CPP-027 | File Repair | Tools extracting properties of the *File* or *Representation* are useful (and sometimes even necessary) for identifying erroneous format structures. |
| Affects | CPP-018 | Community Watch | Either due to changing significant properties or due to updated tools, metadata extraction requirements can be affected. |

# 3. Links to frameworks

## Certification

| Certification framework | Term used in framework to refer to the CPP | Section |
|---|---|---|
| CTS<br>Link | “quality control checks” | Section R10 (Quality assurance) implicitly requires metadata extraction as one of the processes implementing “quality control checks in place [that] ensure the completeness and understandability of data and metadata”. |
| Nestor Seal<br>Link | “technical metadata collect[ion]” | C30 Technical metadata |
| ISO 16363<br>Link | / | / |

## Other frameworks and reference documents

| Reference | Term used in framework to | Section |
|---|---|---|

| Document | refer to the process | |
|---|---|---|
| OAIS | / | / |
| PREMIS | Characterization<br>Metadata extraction | The PREMIS Data Dictionary mentions this operation as “characterization” (p. 249, section Special Topics / Format information), but the event type vocabulary maintained by the PREMIS Editorial Committee at id.loc.gov uses the term “metadata extraction”. |

# 4. Reference implementations

## Example use case(s)

### Metadata Extraction from AV material

| Institutional Background | |
|---|---|
| Institution | Bibliothèque nationale de France, France |
| Hyperlink | / |
| Description | |
| Trigger event | AV material must be analysed by proper tools, beyond format identification, in particular because format identification generally returns information about the container, while AV *Files* are wrapping streams of different nature and format. |
| Problem statement | Discussion with AV experts required that several quality properties be extracted, in particular properties related to the group of pictures.<br>XML was the preferred syntax for the extractor tool output, as it could be easily wrapped in METS *Files*. |
| Proposed solution | BnF has selected the tool MediaInfo as its extractor tool for AV *Files*, according to requirements collected by BnF. MPEG-7, one of its output formats, being standardised and expressed in XML, was selected as the format for *Metadata* to be stored in Archival Information Packages. As MediaInfo provides natively MPEG-7 as one of its output formats, no mapping from the tool output to a standard metadata format was required. |

## Publicly available documentation

| Institution | Organisation type | Language | Hyperlink |
|---|---|---|---|
| TIB – Leibniz Information Centre for Science and Technology and University Library, Germany | National library | English | https://wiki.tib.eu/confluence/spaces/lza/pages/93608618/Ingest |
| | Non-commercial digital preservation service | | |
| | Research infrastructure | | |
| | Research performing organisation | | |
| CSC – IT Center for Science Ltd., Finland | Non-commercial digital preservation service | English | https://urn.fi/urn:nbn:fi-fe2020100578096 (section 5) |
| Archivematica | Digital preservation system | English | https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/preservation/preservation-planning/#characterization |