# File Normalisation (CPP-026)

| | |
|---|---|
| **CPP-Identifier** | CPP-026 |
| **CPP-Label** | File Normalisation |
| **Author** | Kris Dekeyser |
| **Contributors** | Mikko Laukkanen, Juha Lehtonen |
| **Evaluators** | Matthew Addis, Felix Burger, Maria Benauer |
| **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 performs operations on the data *Objects* prior to ingest in order to comply with its format requirements.

## Inputs and outputs

| Input(s) | |
|---|---|
| Data | *Submission Information Package* |
| | *File* |
| Metadata | *Technical metadata* (Format Identifier) |
| Documentation / guidance | File format policy - Preferred formats |
| Output(s) | |
| Data | New *File* |
| | Updated *Submission Information Package* |
| Metadata | *Provenance metadata* (Normalisation events, original and normalised formats, timestamp) |

## Definition and scope

In the context of digital preservation, File Normalisation refers to the process of converting incoming *Files* into standardised, well-supported formats that are more likely to be usable and accessible in the long term. These formats are captured in the preferred preservation formats policy of the TDA. The goal is to ensure the long-term usability, accessibility, and intelligibility of digital content by avoiding dependence on obsolete or proprietary formats and start the repository with a consistent, manageable and sustainable set of formats from the start.

File Normalisation can be driven by an identified risk, in which case it requires that some formats have been assessed a format risk. For each of these formats at risk an alternative format should be available as well as the necessary tools to convert *Files* from the format at risk to its alternative format.

*File* normalisation policy can also be driven by specific organisational requirements like the availability of specialised applications, viewer preferences (e.g. JPEG2000 for IIIF), or the ability to make subsequent processes easier and more efficient such as creating derivatives, tracking and managing risks and obsolescence, and monitoring and reporting of formats in the archive.

If *Files* are submitted in a format not part of the list of preferred preservation formats and there is no normalisation path available to convert them to a supported format, the outcome is undefined. It is up to the TDA to decide what happens.

It may be that the original *File* is submitted in the *SIP* along with the normalised *File*. In this case, the original *File* can only be in bit-level preservation. Reasons for this may be, for example, to ensure the authenticity of the digital *Object*, because it is anticipated that the original *File* format is expected to be useful for the designated community for a reasonable time.

Normalisation differs from other processes that create new *Files* or *Representations*:

- Unlike Normalisation, **Creation of Derivatives** reproduces only the information and significant properties of the original that are useful to address specific needs of the designated community. Thus, the output of **Creation of Derivatives** may lack part of the information that is not required to satisfy said needs.
- Whereas Normalisation aims at reducing the number of formats by converting *Files* or *Representations* to preservation formats defined by the institutional formats policy, **File Migration** is performed to address a specific preservation risk.

# Process description

## Trigger event(s)

| Trigger event | CPP-identifier |
|---|---|
| Submission of an *Information Package* | CPP-029 (Ingest) |
| Change in file format policy - Preferred formats | CPP-018 (Community Watch) |

## Step-by-step description

| No | Supplier | Input | Steps | Output | Customer |
|---|---|---|---|---|---|
| 1A | CPP-029 (Ingest) | Each *File* in the *IP* | Perform the Format Identification process on the *File(s)* | | |
| 1B | CPP-018 (Community Watch) | Change in file format policy - Preferred formats | Find *Files* that require change of normalisation by format, then process each one | | |
| 2 | CPP-008 (File Format identification) | Detected format(s) | Compare the detected format with the list of preferred formats | Complete match: Process completed (step 6) | |
| | | File format policy - Preferred formats | | Deviation: Determine if the policy provides a normalisation plan (i.e. sequence of tools and configuration parameters to use on the *File*) (step | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 3) | |
| 3 | CPP-012 (Risk Mitigation) | Normalisation path | Execute the normalisation plan | Normalised *File* | |
| 4 | | Validation plan | If present, run the validation plan to determine if the normalised *File* provides a similar rendition like the original *File*. If no validation plan is available, it is up to the TDA's policy to decide if manual inspection is needed. | Validation result: OK (step 5a) | |
| | | | | Validation result: Error (5b) | |
| 5a | | OK | Replace the original *File* with the normalised *File* (or add it) in the *SIP* or *AIP* and adapt the *Fixity metadata* and *Technical metadata* | updated *SIP* or *AIP* | |
| 5b | | Error | Report the error for the system or operator to decide further action | Stop the process | |
| 6 | | | Log the normalisation event and its outcome | *Provenance metadata* in the *SIP* | |

## Rationale(s)[1] and worst case(s)

| Rational | Impact of inaction or failure of the process |
|---|---|
| Store information in supported formats | Format migration at a later point could be impossible |
| Anticipate on formats becoming less popular | Attempt to migrate data when the tools to do so are no longer available or knowledge about the format is lost |
| Reducing complexity and risk by normalising to a smaller set of well-chosen formats, TDAs significantly reduce the technical overhead and resources needed for preservation activities. | TDA must maintain too many file formats, which each in turn needs specific knowledge, tools and ongoing maintenance to ensure continued access |

# 2. Dependencies and relationships with other CPPs

## Dependencies

| CPP-ID | CPP-Title | Relationship description |
|---|---|---|
| CPP-008 | File Format Identification | The format information of the *File* in question is one of the deciding input parameters when considering File normalisation. |
| CPP-012 | Risk Mitigation | CPP-012 is in charge of designing normalisation paths whose output would retain all significant properties. |
| CPP-022 | Significant Properties Definition | Like Format migration, File Normalisation should be evaluated based on significant properties. |
| CPP-023 | Risk definition and extraction | The level of risk of the detected format is one of the factors that drives the preferred preservation formats policy. |
| CPP-005 | Identifier Management | Soft dependency (i.e. may require): A normalised file format may be assigned with a new PID. |

[1] Term derived from PREMIS.

## Other relations

| Relation | CPP-ID | CPP-Title | Relationship description |
|---|---|---|---|
| May be required by | CPP-029 | Ingest | The ingest may require that the digital *Objects* are first normalised before ingestion. |
| Facilitates | CPP-015 | Emulation and Rendering Tools | Restricting the formats to those that are well supported by the tools improves the success of this process. |
| Affinity with | CPP-014 | File Migration | Normalisation is performed at Ingest and aims at reducing the number of formats preserved by converting *Files* or *Representations* to preservation formats defined by the institutional formats policy, while Format migration is performed after ingest to address a specific preservation risk. |
| Not to be confused with | CPP-028 | Creation of Derivatives | Unlike Normalisation, Creation of Derivatives reproduces only the information and significant properties of the original that is useful to address specific needs of the Designated Community. Thus, the output of Creation of Derivatives may lack part of the information that is not required to satisfy said needs. Normalisation aims to maintain all the significant properties of the original. |

# 3. Links to frameworks

## Certification

| Certification framework | Term used in framework to refer to the CPP | Section |
|---|---|---|
| CTS<br>Link | normalising | page 7 of 19<br>Levels of Curation: it regards normalisation a requirement to achieve Curation Levels above B. |
| Nestor Seal<br>Link | normalisation | page 34<br>normalisation mentioned as required to document for c17 Authenticity: Ingest |

| | | |
|---|---|---|
| ISO 16363<br>Link | normalisation mentioned in examples | 4.1.1 - preservation objectives<br>4.2.2 - describing transformations<br>4.2.3.3 - provenance information |

## Other frameworks and reference documents

| Reference Document | Term used in framework to refer to the process | Section |
|---|---|---|
| OAIS<br>Link | file format conversion | OAIS mentions in the Generate API function of the Ingest Functional Entity: ‘may involve file format conversions’ page 4-7 |
| PREMIS<br>Link | Normalization | Glossary, page 271:<br>Normalization: Form of Migration in which a version of a Digital Object is created in a new Format with properties more conducive to preservation treatment. Normalization is often implemented as part of the Ingest process. |

# 4. Reference implementations

## Publicly available documentation

| Institution | Organisation type | Language | Hyperlink |
|---|---|---|---|
| CSC – IT Center for Science Ltd., Finland | Non-commercial digital preservation service | English | https://urn.fi/urn:nbn:fi-fe2025040925236 (section 7.7) |
| Archivematica | Digital preservation system | English | https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/ingest/ingest/#normalize ; and https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/preservation/preservation-planning/#normalization<br>Tableau1_3 |