

File Migration (CPP-014)

CPP-Identifier	CPP-014
CPP-Label	File Migration
Author	Bertrand Caron
Contributors	Kris Dekeyser
Evaluators	Matthew Addis, Maria Benauer, Felix Burger
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 supports batch modifications of previously ingested *Files* to prevent a preservation- or access-related risk.

Inputs and outputs

Input(s)	
Data	Source <i>File</i> or Representation whose property(ies) are(were) flagged as a risk that requires action.
Metadata	<i>Technical Metadata</i> (source <i>File</i> or <i>Representation</i>)
Documentation / guidance	Preservation action registry
Output(s)	
Data	Target <i>File</i> or Representation which successfully passes the requirements of risk assessment.
Metadata	<i>Technical Metadata</i> (target <i>File</i> or <i>Representation</i>)
	<i>Provenance metadata</i>

Definition and scope

File migration is the production of one or several new *Representation(s)* in response to a risk assessment, with the goal of replacing the original *File* or *Representation*¹ with an equivalent containing the same information and supporting the same features.

File migration requires careful handling of both the main stream as well as additional content and internal *Metadata*. As many migration tools are only capable of converting the main stream, *Metadata* may need separate processing. For such cases, the use of additional tools for migrating *Metadata* (such as Exiftool) is required to ensure the completeness of the new *Representation*.

Format Migration may target not only the (container) file format but also any property of the *Files* or *Representations* that were considered at risk by **Risk Definition and Extraction** (CPP-023). For example, a TDA may decide to perform Format migration to decrypt protected PDFs as encryption is identified as a preservation risk. For another example see the use case “4-channel JPEG” in CPP-023.

¹ The choice to actually retain the source *Representation* or not is then up to the Archive. Cf. OAIS v. 3, p. 5-7 “At the Archive’s discretion, that first version may be retained for verification of information preservation.”

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

- Unlike **Creation of Derivatives** (CPP-028), Format migration reproduces the information and characteristics of the original to create a new *Representation* that will act as a new authoritative preservation copy. Thus, the output of format migration must contain all information from the original that is deemed significant by the TDA, and ideally most other information as well.
- Whereas **Normalisation** (CPP-026) is performed during Ingest and aims at reducing the number of formats by converting *Files* or *Representations* to preservation formats defined by the institutional formats policy, Format migration is performed after ingest to address a specific preservation risk.

Despite its common usage, the term 'migration' is labeled differently across established frameworks. The definition of "File Migration" used in this CPP is a close match to "Format / forward migration" as used in PREMIS. Its corresponding OAIS notion is "transformation", defined as "*a Digital Migration in which there is an alteration to the Content Information or PDI of an Archival Information Package*".

According to current good practice in digital preservation, format migration after ingest is a much less common operation than anticipated in the early years of digital preservation. Notwithstanding this, it remains important to regard it as a core process as some special events might require its performance at scale.

Process description

Trigger event(s)

Trigger event	CPP-identifier
Alert about risks to the readability, understandability or usability of <i>Files</i> - related to their format or to some other property they bear	CPP-023 (Risk Definition and Extraction)

Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1	CPP-023 (Risk Definition and Extraction)	Properties to be changed and their current value	Identify <i>Files</i> or <i>Representations</i> that share these properties and should therefore be migrated	<i>Files</i> or <i>Representations</i> with risk properties	
		Digital Archive Database			
2	CPP-012 (Risk Mitigation)	Preservation Action Plans	Define target property(ies), based on input from CPP-012	Aspired value of properties (after successful performance of the process)	
3	CPP-012 (Risk Mitigation)	Preservation action registry	Request a migration path	Applicable migration path	
		Properties to be changed and their current value			
4		<i>Files</i> or <i>Representations</i> at risk	Gather a representative test set	Subset	

5		Subset	Perform the selected migration path	Target <i>File</i> or <i>Representation</i>	
		Applicable migration path			
6	CPP-008 (File Format Identification)	Target <i>File</i> or <i>Representation</i>	Apply the characterisation processes against target <i>File</i> or <i>Representation</i> , optionally by running the ingest process and all its subprocesses (if necessary)	Target <i>File</i> or <i>Representation</i> properties	
	CPP-009 (Metadata Extraction)				
	CPP-010 (File Format Validation)				
7		Target <i>File</i> or <i>Representation</i> properties	Control the properties of the target <i>File</i> or <i>Representation</i> against the expected outcome of the process manually or in an automated way, in particular: <ul style="list-style-type: none"> - Properties to be changed by the migration process have the expected value; - Significant properties are maintained; - Target <i>Files</i> or <i>Representations</i> are valid. 	Decision whether the applicable migration path should be confirmed	
	CPP-022 (Significant Properties Definition)	Significant properties			
		Properties expected to change			
	CPP-010 (File Format validation)	Target <i>Files</i> ' validity status			

8		<i>Files or Representations</i> identified at step 2	If the applicable migration path is confirmed, resumption of steps 5 to 7 on the whole set of <i>Files or Representations</i>		
		Confirmed migration path			
9		Confirmed migration path	If the applicable migration path is new to the TDA, record it in the preservation action registry	Updated preservation action registry	CPP-012 (Risk Mitigation)
10		Target <i>File or Representation</i> properties	Decide whether the source <i>File or Representation</i> should be retained		
		Significant properties, properties expected to change			
11			Document in the Provenance Information the creation of a new <i>Representation</i>	<i>Provenance metadata</i> : Documentation of the performed migration date, agents, and properties changed	
11		Target and source <i>File/Representation</i> OR target <i>File/Representation</i> only	Trigger new <i>AIP</i> Version	New <i>AIP</i> Version	CPP-021 (AIP Versioning)
		<i>File or Representation</i> properties			
		<i>Provenance metadata</i>			

Rationale(s)² and worst case(s)

Rational	Impact of inaction or failure of the process
If <i>Files</i> or <i>Representations</i> are identified as at risk in their current format, and the threat is not mitigated through format migration, then the content of <i>Files</i> or <i>Representations</i> may become inaccessible and unusable over time. In the context of FAIR, not migrating to a format that is usable means that data may also lose Interoperability and Reusability, either directly or because <i>DIPs</i> can no longer be created easily for a designated community.	The contents of <i>Files</i> or <i>Representations</i> cannot be used either in whole or in part.

2. Dependencies and relationships with other CPPs

Dependencies

CPP-ID	CPP Title	Relationship description
CPP-005	Identifier Management	Soft dependency (i.e. may require): During file migration, the migrated file format may be assigned a new identifier.
CPP-008	File Format Identification	File format migration requires that the format of the <i>Files</i> is known with certainty.
CPP-009	Metadata Extraction	<i>File</i> format identification is generally limited to an indication of the container format, while migration can apply to any property of the <i>Files</i> . <i>Technical metadata</i> extraction is required to both assess the compliance of files format to the Archive's format policy and control the outcome of the migration.
CPP-010	File Format Validation	Format Validation process should be undertaken after the Format Migration was performed to ensure that the target <i>Files</i> or <i>Representation</i> are valid.
CPP-012	Risk Mitigation	Risk Mitigation determines or selects migration paths.
CPP-022	Significant Properties Definition	Format migration, as it implies the production of a <i>Representation</i> supposed to act as a preservation copy,

² Term derived from PREMIS.

		must rely on significant properties to determine its success or failure.
CPP-023	Risk Definition and Extraction	Risk Definition and Extraction identifies risks related to file format that would trigger a File format migration and provides the method to detect these risks in <i>Files</i> .

Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Required by	CPP-013	Object Management Reporting	File migration provides information on the outcome of the process, tools used.
Facilitates	CPP-015	Emulation and Rendering Tools	May be needed to support rendering tools in the long term.
Affinity with	CPP-026	Normalisation	Normalisation is performed at Ingest and aims at reducing the number of formats preserved by converting <i>Files</i> or <i>Representations</i> to preservation formats defined by the institutional formats policy, while Format migration is performed after ingest to address a specific preservation risk.
Affinity with	CPP-027	File Repair	File repair implies changing the <i>File</i> 's bitstream in order to correct structural issues that could affect reuse and rendering. CPP-027 and Format Migration have thus several steps in common (documentation, control, decision on whether to retain the source <i>File</i> , etc.).
Not to be confused with	CPP-028	Creation of a Derivative	Creation of a derivative creates an additional <i>Representation</i> to be retained, while the outcome of migration and normalisation is supposed to replace its source.

3. Links to frameworks

Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS Link	Format migration	R09 “Preservation Plan”
Nestor Seal Link	Migration	C24 Interpretability of the archival information
ISO 16363 Link	Migration	4.3.1 The repository shall have documented preservation strategies relevant to its holdings.

Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS	Transformation	Definitions in section 1.6.2. Stages when transformation is performed are described in the following sections: <ul style="list-style-type: none">- 4.4.2: Transformation performed by the Producer (normalisation)- 4.4.3: Transformation performed by the Archive by the Ingest functional entity (normalisation)- 4.4.4: Transformation performed by the Archive by the Storage functional entity (migration) “Transformation” description in section 5.2.4.5
PREMIS	(Format / Forward) migration	“Format migration” is defined by PREMIS as a transformation that is performed to address compatibility issues with newer software and hardware.

4. Reference implementations

Example use case(s)

Discontinuation of Finale and Preservation Strategy for MUS(X) *Files*

Institutional Background	
Institution	Bibliothèque nationale de France (BnF), France
Hyperlink	/
Description	
Trigger event	In August 2024, the software company MakeMusic announced that its software Finale would be discontinued. The maintenance was to end one year after the announcement.
Problem statement	The alternative to Finale that was suggested was Dorico , which does not support the native file formats produced by Finale, MUS and MUSX . The last version of Finale provided an export feature to MusicXML version 4.0.
Proposed solution	<p>A risk assessment led BnF to the conclusion that music scores in format MUS & MUSX should be migrated to MusicXML. Only one migration path was identified, as these proprietary formats were supported only by their creation software.</p> <p>In addition to the MusicXML export, a PDF export was decided as another, complementary, <i>Representation</i>. After testing Finale's PDF export, a visual observation showed that the dynamics and interpretation instructions were not printed at their expected location. The PDF export was thus discarded as not preserving this significant property, in favor of a JPEG export followed by a concatenation to PDF, deemed more faithful to the expected visual aspect of the musical score. The migration path was designed as a double export in two different formats, one of which was followed by another merging operation with pdfunite.</p> <p>In this case, the original MUS(X) <i>Representation</i> was retained along with the MusicXML and PDF <i>Representations</i>.</p>

Audio MP3 to WAV Migration and QA Workflow

Institutional Background	
Institution	Statsbiblioteket, Denmark
Hyperlink	https://web.archive.org/web/20130722233410/http://wiki.opf-labs.org/display/SP/SO4+Audio+mp3+to+wav+Migration+and+QA+Workflow
Description	
Trigger event	<p>In 2013, according to its policy, Statsbiblioteket decided to migrate MP3 <i>Files</i> obtained by Danish radio broadcasts to its preferred format for digitised sound preservation, the WAVE uncompressed format.</p> <p><i>Note that this operation should not be considered “recommended” by the authors of this document. Nevertheless, this use case illustrates clearly how a QA process can be applied to migrated Files.</i></p>
Problem statement	<p>Controlling the outcome of the migration should be done by checking the validity of the <i>Files</i>, but also the following significant properties:</p> <ul style="list-style-type: none"> - Internal <i>Metadata</i> located in the header; - Audio signal comparison by evaluating the waveforms similarity.
Proposed solution	<p>The migration was performed by ffmpeg, then the control step was performed in the following way:</p> <ul style="list-style-type: none"> - The validity of the target <i>Files</i> was controlled by JHOVE2; - The AV <i>Metadata</i> extraction tool ffprobe was used to check the properties of the source and target <i>Files</i>; - Finally, a tool called xCorrSound was used to compare the waveforms.

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/93608641/Preservation+Management#PreservationManagement-Migration ; and https://wiki.tib.eu/confluence/spaces/lza/pages/93608961/Significant+Properties
	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-fe2025040925236 (section 7)
Archivematica	Digital preservation system	English	https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/preservation/preservation-planning/#normalization