

# Risk Definition and Extraction (CPP-023)

<b>CPP-Identifier</b>	CPP-023
<b>CPP-Label</b>	Risk Definition and Extraction
<b>Author</b>	Bertrand Caron
<b>Contributors</b>	Kris Dekeyser
<b>Evaluators</b>	Franziska Schwab, Felix Burger, Maria Benauer
<b>Date of edition completed</b>	29.08.2025
<b>Change history</b>	<b>Comments</b>
Version 1.0 - 29.08.2025	Milestone version

# 1. Description of the CPP

The TDA monitors technology evolution, identifies risks and defines detection methods for these risks.

## Inputs and outputs

Input(s)	
Documentation / guidance	Technology alerts
	Community alerts
Output(s)	
Documentation / guidance	Risk inventory
	Risk properties
	Risk properties detection method

## Definition and scope

Problems in the rendering or reuse of *Files* by the TDA's end users may be caused by several reasons. One is the use of particular features of the file format, which are likely to be wrongly interpreted by rendering tools. Preservation actions (e.g. format migration, file repair, emulation and rendering tools, etc.) may then be based on specific properties (below called "risk properties") that go beyond file format or validity status. Some examples of such risk properties are:

- Discontinued support of file format;
- Erroneous file format structures;
- Encryption;
- External dependencies (e.g. non-embedded fonts);
- Embedded *Files*;
- Advanced multimedia features;
- Specific creating application, when said application is known to have created faulty *Files*;
- etc.

Risk Definition and Extraction is therefore a process in the scope of the broader "risk management" activity, that a) identifies risks caused by the *Files*' specific properties and content, and b) defines a method to detect these risks, based on the *Files*' analysis processes (i.e. format identification, metadata extraction and format validation).

Risk Definition and Extraction is also the process responsible for technology watch. It receives updates from **Community Watch** (CPP-018) to evaluate the impact of the risk with regards to the designated community's skills, objectives and means. The two following activities are therefore in scope of this CPP:

- *Risk Identification* involves cataloging potential threats to digital materials, including technological obsolescence (hardware or software becoming unavailable), media

degradation (physical deterioration of storage devices), format obsolescence ( file formats being no longer supported), organisational risks (loss of institutional knowledge or funding), and environmental hazards (disasters, power failures, or security breaches).

- *Risk Assessment* evaluates the likelihood and potential impact of identified risks. This typically involves analysing factors like the criticality of the digital materials, their uniqueness, the stability of their formats, and the resources available for preservation activities.

The step-by-step description below focuses more specifically on the risk extraction part of the process. This activity is closely related to **Risk Mitigation** (CPP-012) that defines which actions can be undertaken to reduce the likelihood or impact of the risk. Like **Risk Mitigation**, Risk Definition and Extraction focuses on a specific subset of risks - in this case, risks related to intrinsic properties of *Files*. Although of major impact, the definition and detection of risks caused by organisational, financial or security issues are not described in the sections below.

## Process description

### Trigger event(s)

Trigger event	CPP-identifier
Community alerts	/
Technology alerts	/

### Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1	CPP-018 (Community Watch)	Information sources for DP (conference papers, specialised literature and journals, vendor announcements, etc.)	Receive community and technology alerts	Risk	
	CPP-010 (File Format Validation), CPP-027 (File Repair)	Experience gathered by technical analysis			

2		Risk	Evaluate the risk likelihood and impact	Assessment of risk likelihood and impact	
3		Risk	Gather test set of <i>Files</i> (both subject to this risk and free from that risk)	Test set	
4		Risk	Select a candidate tool for identifying the risk	Candidate extractor tool	
5		Test set	Run the candidate extractor tool on the test set	Candidate extractor tool output	
		Candidate extractor tool			
6		Candidate extractor tool output	Determine which property in the candidate extractor tool output helps distinguish those <i>Files</i> that are subject to the risk and those that are not	Successful identification of a risk properties detection method (extractor tool) (step 7)	
				No method identified: resume steps 3-5	
7		Chosen extractor tool output	Determine the interpretation of the extractor tool output	Risk properties detection method: path / expression to extract the risk property from the tool's output <sup>1</sup>	CPP-009 (Metadata Extraction)

---

<sup>1</sup> This interpretation should resolve to a true/false statement about the file being subject to the risk or not. It can be simple (an XPATH to the textual content of an element in an XML output) or more complex. See, about this, the use case about multipage TIFFs below.

## Rationale(s)<sup>2</sup> and worst case(s)

Rationale	Impact of inaction or failure of the process
Risk assessment in general is critical for digital preservation activities. Regarding the subset of risks this process description is focused on, identifying risks related to intrinsic <i>files</i> ' characteristics is required if the TDA aims to do semantic / logical preservation.	The impact of the absence of risk assessment on <i>Files</i> ' characteristics may lead to different problems: <ul style="list-style-type: none"><li>- Broken, undetected dependencies;</li><li>- Unusual features badly supported by rendering tools;</li><li>- etc.</li></ul>

## 2. Dependencies and relationships with other CPPs

### Dependencies

CPP-ID	CPP-Title	Relationship description
CPP-008	File Format Identification	Risks can be related to a file format generic properties and concern all instances of it.
CPP-009	Metadata Extraction	Risks can be related to properties common to one or several file formats.
CPP-010	File Format validation	Risks can be related to specific file format erroneous structures.
CPP-018	Community Watch	Risk is measured against the skills and tools available in the designated community.

### Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Affinity with	CPP-019	Data Quality Assessment	Risk Definition and Extraction and CPP-019 both define properties that the TDA should consider and interpret against the result of CPP-009 (Metadata extraction).

---

<sup>2</sup> Term derived from PREMIS.

Required by	CPP-009	Metadata Extraction	The selection of an appropriate extractor tool depends on requirements from Risk Definition and Extraction.
Required by	CPP-012	Risk Mitigation	Risk mitigation is in charge of assigning mitigation methods to the risks listed in the risk inventory maintained by Risk Definition and Extraction.
Required by	CPP-014	File Migration	Risk Definition and Extraction identifies risks related to file format that would trigger a File Migration and provides the method to detect these risks in <i>Files</i> .

### 3. Links to frameworks

#### Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS <a href="#">Link</a>	/	CTS does not explicitly mention Risk definition and extraction, although it is in the scope of R09 (Preservation planning).
Nestor Seal <a href="#">Link</a>	/	Nestor Seal does not explicitly mention Risk definition and extraction, although C11 Preservation measures are based on a risk definition process.
ISO 16363 <a href="#">Link</a>	"identifying each preservation risk"	ISO 16363 mentions Risk definition and extraction in its section 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 <a href="#">Link</a>	Preservation watch	Risk Definition and Extraction is covered by both Monitor Technology and Preservation Watch functions.

PREMIS <a href="#">Link</a>	/	/
--------------------------------	---	---

## 4. Reference implementations

### Example use cases

#### Identification of PDF preservation risks

Institutional Background	
Institution	KB NL, The Netherlands
Hyperlink	<a href="https://bitsgalore.org/2023/05/25/identification-of-pdf-preservation-risks-with-verapdf-and-jhove.html">https://bitsgalore.org/2023/05/25/identification-of-pdf-preservation-risks-with-verapdf-and-jhove.html</a>
Description	
Trigger event	In this blog post, Johan van der Knijff lists a number of PDF features that represent a risk for the rendering or reuse of the <i>file's</i> content. It also indicates how these can be detected using two of the most widely used PDF metadata extractors among the digital preservation community.
Problem statement	Encryption or password-protection, multimedia content or external dependencies may be the cause of the TDA's inability to give to its end user access to the <i>file's</i> content. Format identification and format validation are not sufficient, as these risks are caused by standard features, though these features are used by only some instances of the file format.
Proposed solution	Van der Knijff provides the reader with a method to determine whether the <i>file</i> is using such a feature, and is therefore subject to the risk. Mitigation of such risks is not described in this blog, and the capacity of the TDA or designated community to implement it depends very much on its skills and means.

#### Multi-page TIFFs

Institutional Background	
Institution	Rosetta Users Group
Hyperlink	<a href="https://docs.google.com/spreadsheets/d/1tSpS_1rCeVgOI0dEk_mRu8cR9VjLIYj_CnehswD1aM0/edit?gid=0#gid=0">https://docs.google.com/spreadsheets/d/1tSpS_1rCeVgOI0dEk_mRu8cR9VjLIYj_CnehswD1aM0/edit?gid=0#gid=0</a>



Description	
Trigger event	TIFF <i>files</i> may contain multiple images. Because this is not a common feature of TIFF <i>files</i> , rendering tools may not handle properly images beyond the first one, as they are allowed to take into account only the “baseline” features described in the TIFF specification and ignore its extensions.
Problem statement	<p>Rosetta is providing its users with a risk report. The cause of the risk can be one of the following: Certificate, Encryption, End of life, Non-Standard, Obsolete, Outdated, Propriety.</p> <p>Rosetta users want to monitor risks and identify <i>Files</i> or <i>Representations</i> that are subject to those risks to be able to take action on all <i>Objects</i> previously ingested.</p> <p>The <a href="#">spreadsheet</a> provides a list of risks and the way to detect these; the detection can be based on the file format identifier (PUID, in this case) or by a property of the <i>file</i>.</p>
Proposed solution	<p>The line titled “TIFF with multiple pages” is a risk property, identified by technology and community alerts. Multiple tools can detect this but JHOVE was preferred as it is widely used for both validation and metadata extraction for TIFFs in most of the Rosetta implementations.</p> <p>Nevertheless, the interpretation of JHOVE’s output is complex; it is described in column D as a set of conditions resolving to TRUE or FALSE.</p>

## 4-channel JPEGs

Institutional Background	
Institution	KU Leuven
Hyperlink	
Description	
Trigger event	End users reported some ingested JPEGs did not display correctly. The JPEGs are rendered as a black rectangle. A number of reports had entered, but without a reference to the faulty JPEGs. At last a report was filed with a PID that contained 4 JPEGs, of which 1 did not render.
Problem statement	By comparing the good and bad JPEG’s <i>Technical metadata</i> , we noticed that the bad JPEG had 4 samples per pixel instead of 3. Color space information was missing, so we first assumed that there was an alpha channel added.

	<p>Further investigation into the specification and internet articles pointed out that an alpha channel was highly unlikely, so we downloaded the files data and more tools against the JPEGs. Exiftool listed a Color Transform value of YCCK, which was confirmed by looking at the file's APP14 block data.</p> <p>Closer inspection of the web viewer showed indeed small variations in the blacks where there was a lot of contrast in the original image. That seemed consistent with a YCbCr interpretation of the YCCK data.</p> <p>We decided to investigate all the JPEG files with 4-channel images and all those images turned out to be indeed in the YCCK colorspace.</p>
Proposed solution	<p>We decided all JPEG files need to be converted to YCbCr, but all tools we tried seem to result in a file with colors slightly different from the originals. Still deciding with the producer on the conversion process.</p>

## Publicly available documentation

Institution	Organisation type	Language	Hyperlink
TIB – Leibniz Information Centre for Science and Technology and University Library, Germany	National library	English	<a href="https://wiki.tib.eu/confluence/spaces/lza/pages/93608641/Preservation+Management#PreservationManagement-Riskmanagement">https://wiki.tib.eu/confluence/spaces/lza/pages/93608641/Preservation+Management#PreservationManagement-Riskmanagement</a>
	Non-commercial digital preservation service		
	Research infrastructure		
	Research performing organisation		
CSC – IT Center for Science Ltd., Finland	Non-commercial digital preservation service	English	<a href="https://urn.fi/urn:nbn:fi-fe2023062157386">https://urn.fi/urn:nbn:fi-fe2023062157386</a> (section 2.1)
Archivematica	Digital preservation system	English	Archivematica can provide some of the information needed to support risk assessment as part of its Transfer process. <a href="https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/#transfer-tab-microservices">https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/#transfer-tab-microservices</a>