

# Data Corruption Management

## CPP-004

<b>CPP-Identifier</b>	CPP-004
<b>CPP-Label</b>	Data Corruption Management
<b>Author</b>	Johan Kylander
<b>Contributors</b>	Bertrand Caron, Juha Lehtonen
<b>Evaluators</b>	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 replaces damaged *Files* from replicated copies and reports on actions taken.

## Inputs and outputs

Input(s)	
Data	<i>AIP(s)</i>
Metadata	<i>Fixity metadata</i> (checksums, algorithms and timestamps)
Documentation / guidance	Copy management policy
Output(s)	
Data	<i>AIP(s)</i>
Metadata	<i>Provenance metadata</i> (event date, details and agents involved)
	<i>Fixity metadata</i> (checksums, algorithms and timestamps)
	<i>Storage Management information</i> (storage location)

## Definition and scope

Data Corruption Management is a process, where a TDA restores corrupted *AIPs* from parallel copies. Corrupted *AIPs* are identified and flagged by the **Integrity Checking** (CPP-003) process, which periodically scans the fixity of *AIPs*.

When **Integrity Checking** (CPP-003) has flagged that an *AIP* is corrupted (i.e. that it has been unintentionally altered), the TDA recovers an intact copy from another storage medium in order to replace the corrupted *AIP*. As part of the retrieval process, the checksums of the copied data are validated to verify a) the integrity of the data that is used as a source and, b) that the new target has been copied successfully. This is similar to the process of **Replication** (CPP-011). Subsequently, *Provenance metadata* is updated to maintain a record of the replacement procedure.

The replacing copy can be written to another storage medium than the corrupted one (e.g. in the case of magnetic tapes, where read-write operations to a single tape are kept to a minimum). In these cases, *Storage management information* must also be updated to reflect the new location of the copy.

The TDA may choose to replace the whole storage medium's content, triggering a process similar to **Refreshment** (CPP-030). This can be done when media-wide corruption or read/write errors are detected.

Data Corruption Management relies on the TDA having several parallel copies, preferably on different storage media and in different storage locations. The number of parallel copies, and their storage conditions are defined in the TDA's policies as maintained by **Risk Mitigation**

(CPP-012) and a copy management policy in particular. In accordance with agreed on best-practices, at least three copies are recommended, as there should exist at least two other valid copies in case one copy is corrupted or destroyed.

## Process description

### Trigger event(s)

Trigger event	CPP-identifier
An <i>AIP</i> or <i>File</i> that has been flagged as corrupt	CPP-003 (Integrity Checking)
Media-wide corruption or read/write errors are detected	

### Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1a	CPP-003 (Integrity Checking)	Integrity checking report	Identify and locate corrupted <i>AIPs</i>	Inventory of corrupted <i>AIPs</i>	
				Storage medium with broken <i>AIPs</i>	
1b		Report of media-wide errors	Identify and locate the <i>AIPs</i> on the corrupted medium	Inventory of corrupted <i>AIPs</i>	
				Storage medium with media-wide errors packages	
2	CPP-012 (Risk Mitigation)		Select source medium to copy the <i>AIPs</i> from	Authoritative storage medium for replicating/copying the <i>AIPs</i>	

3			Select target storage medium (can, and often is, be same as the original storage medium identified in step 1a)	Target storage medium	
3b			In case of media-wide errors: Provision of a fresh storage medium that will replace the old one	The fresh storage medium that will replace the old one	
4		Source storage medium of <i>AIPs</i>	For each <i>AIP</i> individually, start the copy process (steps 5 to 9):		
		Target storage medium			
		Inventory of <i>AIPs</i> involved in the process			
5			Retrieve the <i>AIP</i> from the source storage medium		
6			Copy the <i>AIP</i> to the target storage medium	New copy of <i>AIP</i>	
7		Existing <i>Fixity metadata</i>	Validate the fixity of the <i>AIP</i> on the fresh storage medium	Valid status (step 8)	<i>Fixity Metadata</i>
				Invalid status (go back to step 5)	

8			Update the fixity for the new <i>AIP</i> copy	<i>Fixity metadata</i>	
9			If the target storage medium is different from the original storage medium with the broken <i>AIP</i> :  Update the storage location for the new <i>AIP</i> copy	<i>Storage management information</i>	
10		Inventory of <i>AIPs</i> involved in the process	Check that all <i>AIPs</i> in the inventory have been successfully copied	Confirm completeness of the copy process (step 11)	
				Error (go back to copy process loop)	
11			In case of media-wide errors:  Update information about the fresh storage medium (e.g. <i>File</i> locations, media identifiers) and mark the old medium and its contents as ready for deletion/decommissioning	<i>Storage management information</i>	
12			Document the event and its timestamp	<i>Provenance Metadata</i>	
13		Original storage medium that has been refreshed	In case of media-wide errors:  Ensure data security and that confidentiality is not compromised by making sure that data on the		

			original storage medium is properly deleted		
14			Decommission the old storage medium	Record of decommissioning	

## Rationale(s)<sup>1</sup> and worst case(s)

Rationale	Impact of inaction or failure of the process
Parallel copies of the data	Corrupted data cannot be restored unless parallel copies exist.
<i>Fixity metadata</i>	Data corruption cannot be detected, and replaced copies cannot be verified, without <i>Fixity metadata</i> .

## 2. Dependencies and relationships with other CPPs

### Dependencies

CPP-ID	CPP-Title	Relationship description
CPP-003	Integrity Checking	Corrupted data is detected by periodic integrity checking.
CPP-005	Identifier Management	Soft dependency (i.e. may require): If a <i>File</i> is corrupted, it may need to be repaired or replaced. During this process, a new <i>PID</i> may be created.
CPP-012	Risk Mitigation	The number of parallel copies and how they are stored (media, locations) are defined in a <i>TDA's</i> policy that arises out of mitigating risks to preserved data.

### Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Required by	CPP-013	Object Management Reporting	Fixing corrupted <i>AIPs</i> produces <i>Provenance metadata</i> and data for quality reporting to the stakeholders.
Affinity with	CPP-002	Checksum validation	All new <i>AIP</i> copies must have their checksum validated to verify that the process was successful. The checksum validation is more mechanical in its

---

<sup>1</sup> Term derived from PREMIS.



			nature in Data Corruption Management, only aiming at verification of the copy process. In contrast to CPP-002, it does not have to negotiate with producers or examine the results.
Affinity with	CPP-011	Replication	Corrupted copies are replaced by intact copies, effectively replicating the intact copy, but not creating a new parallel copy.
Not to be confused with	CPP-007	Virus Scanning	If a <i>File</i> is detected as infected and cannot be cleaned, it might be considered "damaged." However, CPP-004 typically applies to technical corruption or loss, rather than deliberately human-made damage such as malware-infected <i>Files</i> (CPP-007). In practice, infected <i>File</i> are more likely to be replaced (by the producer) or rejected.
Triggers	CPP-017	Disposal	Data corruption management may trigger the disposal.
Triggers	CPP-030	Refreshment	Media-wide corruption triggers a refreshment process where the data to be copied is not retrieved from the corrupted medium, but an intact one.
Alternative to	CPP-027	File Repair	File Repair is an alternative (fallback) to Data Corruption Management in cases where no intact copy of corrupted or broken data is available, since repairing the structure of an altered copy is the only option.

### 3. Links to frameworks

#### Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS <a href="#">Link</a>	"For each storage location, measures should be in place to ensure that unintentional or unauthorised changes can be detected and correct versions	R14 Storage & Integrity

	of data and metadata recovered”	
Nestor Seal <a href="#">Link</a>	“restoration of the archival information packages” “recovering archival information packages in the event of damage”	C15 Integrity: Functions of the archival storage
ISO 16363 <a href="#">Link</a>	“Recovery actions”	5.1.1.3.1 The repository shall record and report to its administration all incidents of data corruption or loss, and steps shall be taken to repair/replace corrupt or lost data.

## Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS <a href="#">Link</a>	Disaster Recovery	4.2.3.4
PREMIS <a href="#">Link</a>	/	/

## 4. Reference implementations

### Example use case(s)

#### Summary between 2021-2024 at CSC

Institutional Background	
Institution	CSC – IT Center for Science Ltd., Finland
Hyperlink	<a href="https://digitalpreservation.fi/en/services/quality_reports">https://digitalpreservation.fi/en/services/quality_reports</a>
Description	
Trigger event	2024a: Software error 2024b: Scheduled Integrity Checking 2022a: Human error 2022b: Scheduled Integrity Checking 2021: Scheduled Integrity Checking
Problem statement	2024a: Contents of one tape were lost

	<p>2024b: One corrupt AIP copy on a tape</p> <p>2022a: One corrupt AIP copy on a disk</p> <p>2022b: One corrupt AIP copy on a disk</p> <p>2021: Ten corrupt AIP copies on a tape</p>
Proposed solution	<p>2024a: The tape was restored from other copies</p> <p>2024b: A new copy was produced</p> <p>2022a: A new copy was produced from the tape</p> <p>2022b: The corrupted copy of the package had been unsuccessfully copied to storage during ingest earlier but had been later successfully copied to disk storage automatically.</p> <p>2021: New copies were produced</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/93608373/Archival+Storage#ArchivalStorage-Recovery">https://wiki.tib.eu/confluence/spaces/lza/pages/93608373/Archival+Storage#ArchivalStorage-Recovery</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://digitalpreservation.fi/en/services/quality_reports/2024">https://digitalpreservation.fi/en/services/quality_reports/2024</a> (section Quality Deviations Relating to Preserved Content in 2024)
			<a href="https://digitalpreservation.fi/en/services/quality_reports/2022">https://digitalpreservation.fi/en/services/quality_reports/2022</a> (Quality Deviations Related to the Data in Preservation in 2022)
Archivematica	Digital preservation system	English	<a href="https://www.archivematica.org/en/docs/storage-service-0.23/recovery/#recovery">https://www.archivematica.org/en/docs/storage-service-0.23/recovery/#recovery</a>