

# Ingest (CPP-029)

<b>CPP-Identifier</b>	CPP-029
<b>CPP-Label</b>	Ingest
<b>Author</b>	Mikko Laukkanen, Johan Kylander
<b>Contributors</b>	Bertrand Caron, Mattias Levlin
<b>Evaluators</b>	Kris Dekeyser, Maria Benauer, Felix Burger
<b>Date of edition completed</b>	29.08.2025
<b>Change history</b>	<b>Comments</b>
Version 0.1 - 31.07.2025	Initial version

# 1. Description of the CPP

The TDA performs all operations necessary to transform an *SIP* into an *AIP*.

## Inputs and outputs

Input(s)		
Data	<i>SIP</i>	
Metadata	<i>Fixity metadata</i>	
	<i>Descriptive metadata</i>	
	Optional	<i>Provenance metadata</i>
		<i>Technical metadata</i>
		<i>Rights metadata</i>
Documentation / guidance	Packaging policy	
Data	<i>AIP</i>	
Metadata	<i>Technical metadata</i>	
	<i>Provenance metadata</i>	
	<i>Rights metadata</i>	

## Definition and scope

Ingest is a high-level CPP, that is composed of or utilised by many other CPPs, and refers to the process of acquiring and incorporating data into a TDA. In its most abstract form, Ingest describes the process in which a TDA receives a *SIP* and transforms it into one or several *AIP(s)* through a process that includes **Data Quality Assessment** (CPP-019). At the end of a successful ingest, the data is preserved in the TDA for future use.

The Ingest process begins with the transfer of digital *Objects* and *Metadata* from their source environment to the TDA in the form of a *SIP*. The digital preservation workflow triggered by the submission ensures that the data is properly prepared for long-term storage, discovery and access. The workflow ensures that essential *Metadata* about the creation, structure, and context (CPP-016 **Metadata Ingest and Management**) of the data exists. *Fixity metadata* that will be essential for future preservation actions must also exist before the data can be preserved. In addition, *Technical metadata* that could not be easily extracted (e.g. column delimiters for CSV, quality assessment for OCR, or EPUB *Files* etc.) from the *Files* by the TDA may be required to be supplied by the producer.

**Data Quality Assessment** (CPP-019) measures are integrated throughout the ingest workflow to catch potential issues early. This includes **Virus Scanning** (CPP-007), **File**

**Format Identification** (CPP-008), **Metadata Extraction** (CPP-009) and **File Format Validation** (CPP-010), and completeness checks to ensure that the digital *Objects* are suitable for preservation. The TDA verifies that transferred *Objects* are complete and uncorrupted through **Checksum Validation** (CPP-002). The process also involves assessing whether the digital *Objects* conform to the repository's technical requirements and collection development policies. Depending on the TDA's file format policy - preferred formats, *Objects* may also need to be normalised (CPP-026 **File Normalisation**) to preferred formats before or during the ingest process. During the ingest process, the TDA can generate additional *Metadata* and assign identifiers (CPP-005 **Identifier Management**) to support **Enabling Discovery** (CPP-024); perform **Enabling Access** (CPP-025) for cataloging purposes; and generate *Provenance metadata* that documents the transfer and processing history of the *Objects*. In order to evaluate the *SIP* and its contents, the TDA ensures that the *SIP* structure is valid and that the *SIP* is not incomplete (i.e. all *Objects* and *Metadata* are present).

*SIPs* that conform to the TDAs requirements and policies are transformed into *AIPs* which are sent to preservation in the archival storage of a TDA. *SIPs* that do not conform to the TDAs requirements will be handled according to its policies (in particular, file format policy and validation policy). The TDA can either reject the submitted data, ask the producer to address the issues before proceeding, flag the data as problematic and ingest it as it is, or perform an operation to address the identified issues.

## Process description

### Trigger event(s)

Trigger event	CPP-identifier
Submission of data to a TDA	/

### Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1	CPP-008 (File Format Identification), CPP-009 (Metadata Extraction), CPP-005 (Identifier Management)	Digital <i>Objects</i>	Pre-ingest actions (normalisation, metadata generation, identifier generation etc.) and <i>SIP</i> creation performed by the producer	<i>SIP</i>	CPP-005 (Identifier management)
		<i>Metadata</i> provided by the producer			
2		<i>SIP</i>	Submission of data to a TDA ( <i>SIP</i> transfer)		
3		<i>SIP</i>	Identify whether the <i>SIP</i> is meant to create a new <i>AIP</i> or is intended to update one or several <i>AIPs</i>	Updating request	CPP-021 ( <i>AIP</i> Versioning)

			E.g. If the <i>SIP</i> has a Producer identifier that already corresponds to an <i>AIP</i> ingested in the system, proceed with the process		
		<i>SIP</i>	Ensure that the <i>SIP</i> structure conforms to the requirements and that its contents are not missing	Valid and complete <i>SIP</i>	
		Packaging policy			
4	CPP-002 (Checksum Validation)	<i>Files</i> in the <i>SIP</i>	Perform checksum validation on each <i>File</i> in the <i>SIP</i>	<i>Information package</i> with fixity checked	
5	CPP-019 (Data Quality Assessment)	Quality assessment report	Quality assessment (ensuring that the submitted data conforms to requirements set by the TDA)	<i>Information package</i> with assessed quality	
		<i>SIP</i>		Result of the quality assessment recorded as <i>Provenance metadata</i>	
6a	CPP-008 (File Format Identification), CPP-009 (Metadata Extraction), CPP-007 (Virus Scanning)	<i>SIP</i>	Perform File Format Identification, Metadata Extraction, and Virus Scanning	<i>Technical metadata</i>	
				<i>Provenance metadata</i>	
6b	CPP-010 (File Format Validation)	Format policy - Validation	Optional: Perform Format Validation if the TDAs format policy states that validation must be performed	<i>Technical metadata</i>	
		<i>Files</i> in the <i>SIP</i>		<i>Provenance metadata</i>	

7	CPP-020 (Rights Management)	<i>Objects in the SIP</i>	Perform rights assessment on <i>Objects</i> contained in the <i>SIP</i>	<i>Rights metadata</i>	
		Rights assessment			
8	CPP-016 (Metadata Ingest and Management)	<i>Metadata</i> provided by the producer	Record the <i>Metadata</i> provided by the producer and produced by the TDA according to the TDAs <i>SIP</i> requirements and policy of automatic enrichment of <i>SIP Metadata</i>	<i>Information package with Metadata</i> recorded	
		<i>Technical metadata</i>			
		<i>Provenance metadata</i>			
		<i>Rights metadata</i>			
9	CPP-005 (Identifier Management)	Identifier	Assign identifier to the <i>Information package</i>	<i>Information package</i> with Identifier assigned	
10	CPP-026 (File Normalisation)	<i>Files in the SIP</i>	Optional (only if the TDA supports normalisation during ingest): Normalisation of data, including documenting the actions	<i>New Representations</i> in a supported format	
				<i>Provenance metadata</i>	
11	CPP-028 (Creation of Derivatives)	<i>Files in the SIP</i>	(only if the TDA supports creating derivatives during ingest): Generation of derivatives Optional	New additional <i>Representations</i>	
12 a		<i>Information package</i> with fixity checked, identifier assigned, quality assessed, <i>Metadata</i> recorded and optionally new <i>Representations</i> added	If <i>SIP</i> conforms to the requirements (steps 7a and 7b):  - <i>SIP</i> transformation to <i>AIP</i>	<i>AIP</i>	

12 b		<i>AIP</i>	Move the <i>AIP</i> to the archival storage	<i>AIP</i> on multiple locations	CPP-011 (Replication)
13		Error-handling policies (in particular file format policy and validation policy)	If <i>SIP</i> doesn't conform to the requirements, perform one of these actions <ul style="list-style-type: none"> <li>- Rejection of <i>SIP</i></li> <li>- Request the producer to address the issues</li> <li>- Flag the data as problematic and ingest it as it is</li> <li>- Perform an operation to address the identified issues.</li> </ul>	Error report to the submitter of the <i>SIP</i> (producer)	
				Trigger technical analysis	
14			Notification/report to the producer about the outcome of the ingest	Ingest report	

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

Rationale	Impact of inaction or failure of the process
Ingest transfers the responsibility from the creator/owner/depositor of the digital <i>Objects</i> to the TDA, enabling long-term preservation, discovery and access to the digital <i>Objects</i> . Also, the ingest process captures the <i>Objects</i> ' state at the time of transfer through checksums, metadata extraction, and documentation of the transfer process itself. This creates an auditable trail that supports future authenticity claims and helps detect any corruption or unauthorised modifications that may occur over time.	The digital <i>Objects</i> remain vulnerable to loss, corruption, or unauthorised changes in their original environment. Furthermore, <i>Objects</i> in their original environments are often stored in formats, structures, or contexts that are not optimal for long-term preservation.

## 2. Dependencies and relationships with other CPPs

### Dependencies

CPP- ID	CPP-Title	Relationship description
CPP-002	Checksum Validation	All of these processes must be performed during ingest.
CPP-005	Identifier management	
CPP-007	Virus Scanning	
CPP-008	File Format Identification	
CPP-009	Metadata Extraction	
CPP-020	Rights Management	Some minimal rights assessment must be performed during ingest to verify that the TDA should be in charge of preserving the content of the <i>SIP</i> .
CPP-016	Metadata Ingest and Management	The ingest process produces <i>Technical</i> , <i>Rights</i> and <i>Provenance metadata</i> that are recorded in the <i>Information package</i> and digital archive database by Metadata Ingest

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



		and Management.
CPP-010	File Format Validation	Soft dependency (i.e. may require): A TDA may validate the format of the submitted <i>Files</i> in the ingest phase.
CPP-019	Data Quality Assessment	Soft dependency (i.e. may require): The TDA may have quality requirements that may be checked during ingest.
CPP-026	File Normalisation	Soft dependency (i.e. may require): The ingest may require that the digital <i>Objects</i> are first normalised before ingestion.

## Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Required by	CPP-021	AIP Versioning	Versioning implies several delicate operations, in particular in the case of a partial update, where the incoming <i>SIP</i> should be merged with the existing <i>AIP</i> .
May be required by	CPP-028	Creation of Derivatives	The ingestion may generate derivatives for access.
Affinity with	CPP-012	Risk Mitigation	The ingest process must adhere to the risk mitigation policies.
Affinity with	CPP-013	Object Management Reporting	Ingest is both an important provider of reporting data to the TDA (via other CPPs) as well as a customer, as the ingest checks and outcomes must be reported to the producer.
Triggers	CPP-001	Checksum Generation	A new <i>SIP</i> being submitted and processed triggers Checksum Generation.
Triggers	CPP-002	Checksum Validation	<i>SIP</i> ingest triggers Checksum Validation.

## 3. Links to frameworks

### Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS	ingest	/

<a href="#">Link</a>		
Nestor Seal <a href="#">Link</a>	Ingest	C14 Integrity: Ingest Interface C17 Authenticity: Ingest
ISO 16363 <a href="#">Link</a>	Ingest	4.1. Ingest: Acquisition of content 4.2. Ingest: Creation of the AIP

## Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS <a href="#">Link</a>	Ingest Ingest Functional Entity	4.2.2. General - Figure 4-1 4.2.3.3. Ingest
PREMIS <a href="#">Link</a>	Ingest	Glossary

## 4. Reference implementations

### Example use case(s)

#### Ingest processes described as workflows

Institutional Background	
Institution	<i>Several institutions</i>
Hyperlink	<a href="https://coptr.digipres.org/index.php/Workflow:Community_Owned_Workflows">https://coptr.digipres.org/index.php/Workflow:Community_Owned_Workflows</a>
Description	
Trigger event	Several institutions have described their workflow for ingesting <i>Objects</i> in the Community Owned Workflow section of the COPTR registry.
Problem statement	
Proposed solution	

## 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/93608618/Ingest">https://wiki.tib.eu/confluence/spaces/lza/pages/93608618/Ingest</a>
	Non-commercial digital preservation service		
	Research infrastructure		
	Research performing organisation		
CSC – IT Center for Science Ltd., Finland	Non-commercial digital preservation service	Finnish	<a href="https://urn.fi/urn:nbn:fi-fe2024051731943">https://urn.fi/urn:nbn:fi-fe2024051731943</a> (Appendix 4, section 2.2.1)
Archivematica	Digital preservation system	English	Transfer (steps that lead up to creating a SIP, e.g. checksum generation, file format identification etc.): <a href="https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/">https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/</a> ; Ingest (steps from SIP to AIP, e.g. file format normalisation): <a href="https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/ingest/ingest/">https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/ingest/ingest/</a>