

Checksum Generation and Recording (CPP-001)

CPP-Identifier	CPP-001
CPP-Label	Checksum Generation and Recording
Author	Kris Dekeyser
Contributors	Johan Kylander, Bertrand Caron
Evaluators	Felix Burger, Maria Benauer, Fen Zhang
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 (Trustworthy Digital Archive) records checksums for every *File*.

Inputs and outputs

Input(s)	
Data	<i>File</i>
Documentation / guidance	Storage management policy - Checksum algorithms
Output(s)	
Metadata	<i>Fixity metadata</i> (one or multiple checksum(s) for the <i>File</i> as well as their associated algorithms)
	<i>Provenance metadata</i> (a timestamp or event that describes the checksum calculation)

Definition and scope

A checksum or message digest is a fixed size stream of data generated by a transformation of the *File* data by means of an algorithm. Any change to the data would result in a change to the calculated digest. The algorithm can be a cyclic redundancy check or a cryptographic hash function.

The *File* checksums form an important part of the fixity information which is the cornerstone for performing bit-level digital preservation. The system must keep track of this message digest for each *File*. Due to the possibility of collisions (multiple data streams having the same checksum), storing multiple message digests generated by different algorithms is recommended. The TDA policy should define a list of required algorithms.

The Checksum Generation and Recording process is the action of acquiring and storing the message digests associated with any *File* that the system needs to keep track of. *Files* should come with any number of checksums generated prior to their submission. In that case, the system should store those checksums and use them as-is and new checksums should be generated for those algorithms that are missing.

Process description

Trigger event(s)

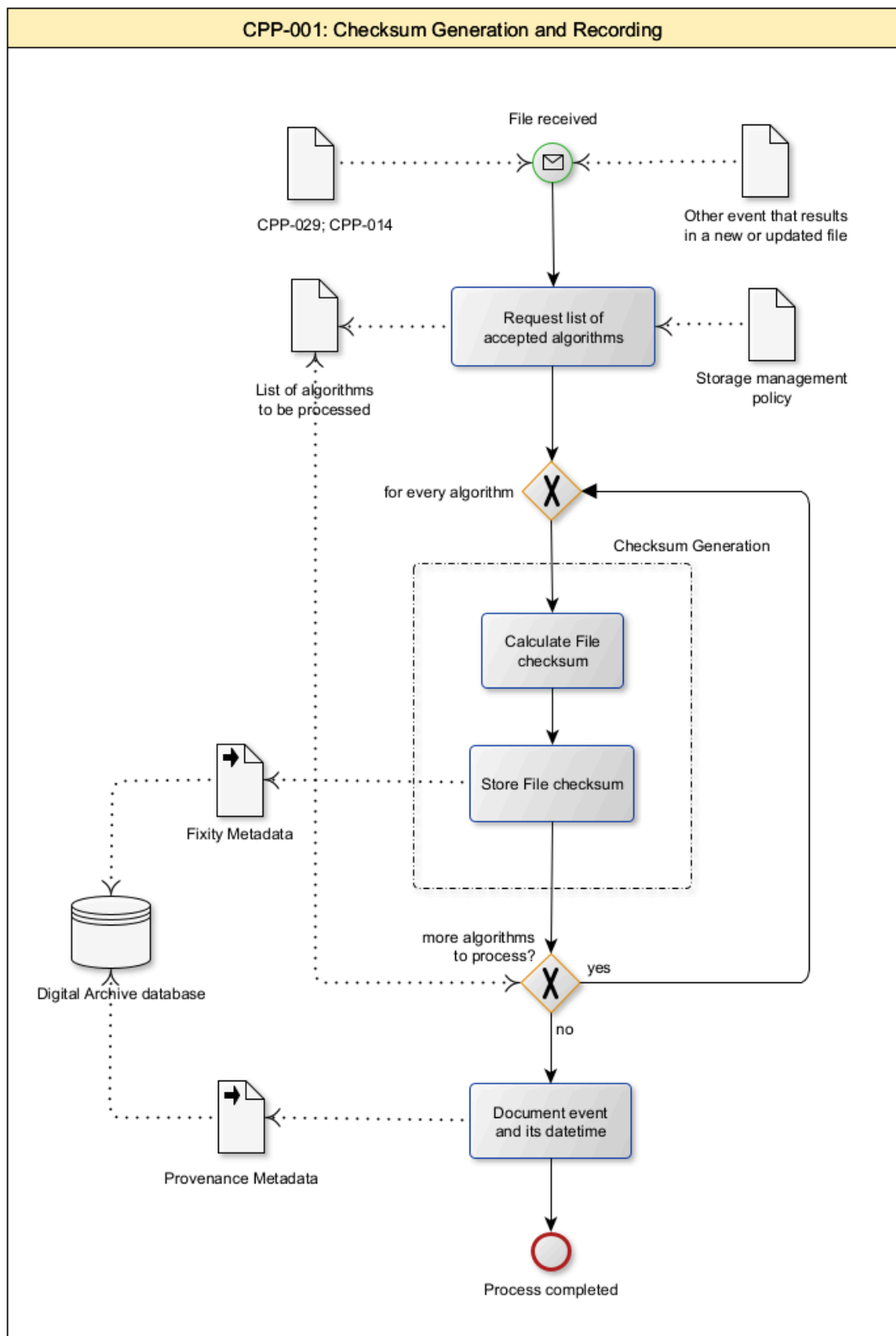
Trigger event	CPP-identifier
A new <i>SIP</i> is submitted and processed	CPP-029 (Ingest)
<i>File</i> update or replacement due to Preservation Action (e.g. migration)	CPP-014 (File Migration)
Any other action that results in a new or updated <i>File</i> being added to the system	

Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1		Storage management policy - Checksum algorithms	Get the list of accepted checksum algorithms	List of accepted checksum algorithms	
2		<i>File</i> List of checksum algorithms	Calculate the checksum for each algorithm	List of checksums for the <i>File</i> based on different algorithms	
3		List of checksums for the <i>File</i> based on different algorithms	Store the checksums in the <i>Fixity metadata</i> for the <i>File</i>	Updated <i>Fixity metadata</i> of the <i>File</i> in the TDA database	CPP-002 (Checksum Validation)

		<i>File</i>			CPP-003 (Integrity Checking)
4		<i>File</i>	Document the event and its datetime	Datetime for the checksum generation and other related <i>Provenance metadata</i>	

BPMN-diagram V1.0



Rationale(s)¹ and worst case(s)

Rationale	Impact of inaction or failure of the process
Keeping track of the fixity information of each <i>File</i>	Corrupted data can get undetected or be detected when it is too late to take corrective action
Event datetime for checksum generation	No starting point from when the fixity can be checked (and guaranteed)
Multiple checksums for each <i>File</i>	Collisions where changes to a <i>File</i> produce the same checksum, are more likely with a single checksum algorithm than with multiple checksum algorithms

2. Dependencies and relationships with other CPPs

Dependencies

CPP-ID	CPP-Title	Relationship description
/	/	/

Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Required by	CPP-002	Checksum Validation	CPP-002 relies on fixity information as produced and stored by CPP-001, when triggered by CPP-025 Enabling Access and CPP-006 AIP Batch Export. When triggered by CPP-029 Ingest CPP-002 rather relies on the fixity information supplied in the SIP.
Required by	CPP-003	Integrity checking	The integrity checking process relies on the fixity information as produced and stored by CPP-001.
Required by	CPP-006	AIP Batch Export	<i>Fixity metadata</i> is used to verify the integrity of data written into the exported AIP.

¹ Term derived from PREMIS.

Required by	CPP-016	Metadata Ingest and Management	The checksums and associated algorithms need to be stored in the <i>File's Fixity metadata</i> .
-------------	---------	--------------------------------	--

3. Links to frameworks

Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS Link	Checksum (cf Extended Guidance documentation)	/Information Technology & Security/Storage & Integrity (R14)
Nestor Seal Link	integrity	C14 Integrity: Ingest Interface
ISO 16363 Link	integrity measurements	Checksum Generation: 3.3.5 The repository shall define, collect, track, and appropriately provide its information integrity measurements.
		Recording: 4.1.6 The repository shall obtain sufficient control over the <i>Digital Objects</i> to preserve them.

Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS Link	/	OAIS does not describe the process of checksum generation and recording in its functional model but defines fixity information as part of <i>Administrative metadata</i> , which is “necessary for adequate preservation of the Content Data Object”.
PREMIS Link	Message digest calculation	The term “message digest calculation” is referenced in the glossary. The general topic of fixity is addressed in section Fixity, Integrity, Authenticity, p. 258.

4. Reference implementations

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/93608951/Metadata#Metadata-TMDTechnicalmetadata
	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-fe2020100578094 (section 2.4.4.2)
Archivematica	Digital preservation system	English	https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/#transfer-tab-microservices