Data Quality Assessment (CPP-019)

CPP-Identifier	CPP-019
CPP-Label	Data quality assessment
Author	Mikko Laukkanen, Juha Lehtonen
Contributors	Bertrand Caron, Johan Kylander
Evaluators	Franziska Schwab, Felix Burger, Maria Benauer
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 evaluates and re-evaluates the data quality of *Information Objects*.

Inputs and outputs

Input(s)	
Data	*Information object*
	File
Metadata	*Descriptive metadata*
	Technical metadata
	Provenance metadata
	Rights metadata
	Structural metadata
Documentation / guidance	Quality assessment policy
	Format policy - preferred formats
	Collection development policy
	Metadata recording policy
Output(s)	
Documentation / guidance	Quality assessment report

Definition and scope

Data Quality Assessment refers to the systematic evaluation of *Objects* and their associated *Metadata* against predefined measures to ensure they meet the standards necessary for consumers' needs and continued access. The assessment typically covers several key dimensions, some of these are for example:

- *Authenticity:* The *Object* is what it purports to be (i.e. it has been created, modified and sent by the person purported to have done it at the date and time purported). The designated community must be able to trust that the data is real and credible and is managed by a trustworthy TDA. Sufficient information must exist to understand the *Object's* creation circumstances, provenance, and relationship to other content. In addition to integrity checks, the authenticity of the data is ensured by controlled changes through preservation actions and the *Provenance metadata*.

- *Completeness:* The *Object* and the *Metadata* are complete. They do not have missing parts or links to targets outside the preserved *Object* which should remain accessible.
- *Consistency***:** The *Object* is presented in applicable file formats or *Representations* with applicable metadata formats. Conflicting values in the *Metadata* should be avoided.
- *Relevance:* The data preservation is based on a predefined collection development policy (i.e. has a purpose of being preserved).
- *Structured:* The structure of the *Object* is described in the *Metadata*. Complex *Objects* are organised, including relationships between *Files*, proper sequencing of multi-part *Objects*, and the integrity of any embedded *Metadata* or links.
- *Understandability:* The information is understandable and meaningful for the designated community.
- *Validity:* The *Object* and *Metadata* are valid against the *File* and metadata format specifications and standards, and comply with all other predefined profiles and rules.

Data quality assessment may include various processes, repeated from time to time. A very common phase to perform an assessment is in the *Ingest* phase, but the use case described below demonstrates that such a process can also be performed at the access stage. The data may be rejected from digital preservation, if it does not meet the criteria. An assessment typically has the following steps, described on a very high level:

1. Define the scope of the assessment;
2. Define data quality dimensions and metrics, including possible thresholds;
3. Gather and analyse data;
4. Create a quality report about all the findings;
5. If needed, update the data and *Metadata* to improve the quality.

The step-by-step description below mainly concentrates on the technical aspects of the data and *Metadata*, but the scope of this CPP is indeed covering a broader range of contextual data quality properties.

The assessment process often employs both automated tools and manual review. For example, automated tools can perform file format identification, validate *File* or *Information package* structures, check for malware, verify checksums, or check for completeness of a delivery against an inventory. Human reviewers, for example, may evaluate content accuracy, *Metadata* completeness, and contextual adequacy. The processes should be automated as much as possible for faster processing and to avoid human errors.

Results from Data Quality Assessment affect preservation planning decisions (e.g. what additional *Metadata* needs to be captured). The assessment also establishes baseline quality metrics that can be monitored over time to detect degradation or other changes that might necessitate intervention.

Process description

Trigger event(s)

Trigger event	CPP-identifier
Ingest	CPP-029 (Ingest)
Metadata ingest	CPP-016 (Metadata Ingest and Management)
Mass export of *AIPs* from the TDA	CPP-006 AIP Batch Export
Periodic re-appraisal	/

Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1	CPP-018 (Community Watch)	Preservation objectives	Based on preservation intent as defined by Community Watch, derive quality properties that will be extracted by other CPPs	Quality properties	
2		Quality properties	The TDA receives a defined set of quality properties and determines what data is required to create a quality assessment report. This triggers steps 3 to 8)	Specification of the data required for the assessment.	
3	CPP-008 (File Format	Specification of the data required for the assessment	If quality properties concern file formats:	Technical quality report	

	Identification)	*File* File format identifier Format policy - preferred formats	Assess the file format against the preferred formats policy		
4	CPP-010 (File Format Validation)	Specification of the data required for the assessment *File* Validity status	If quality properties concern the validity of formats: Assess the validity status.	Technical quality report	
5	CPP-009 (Metadata Extraction)	Specification of the data required for the assessment *File / Representation* Quality properties Extracted *Metadata*	If quality properties concern technical qualities or completeness of *Files* or *Representations*: Assess the technical quality and completeness against quality properties	Technical quality report	
6	CPP-016 (Metadata Ingest and Management)	Specification of the data required for the assessment Metadata recording policy *Object* *Metadata*	If quality properties concern metadata quality: Assess the metadata quality.	Metadata ingest report	

7	CPP-007 (Virus Scanning)	Specification of the data required for the assessment	If quality properties concern existence of malware: Scan for malware	Virus scanning report	
		File			
8	CPP-020 (Rights Management)	Specification of the data required for the assessment	If quality properties concern the legal status and authenticity of the *Object*: Assess the legal status of the *Object*	Legal status report	
		Rights metadata			
		Object			
9		File format identifier	Creation of quality assessment report from suppliers	Quality assessment report	
		Validity status			
		Metadata ingest report			
		Technical quality report			
		Virus scanning report			
		Legal status report			
10		Quality assessment report	Assess the quality of an *Object* during specific stages)(e.g. during ingest)		CPP-029 (Ingest)
11			Optional: The quality of an *Object*, *AIP* or *Metadata* can be enhanced or modified based on the quality assessment report. It may run for example some of the following CPPs:		

			● CPP-014 (File Migration) ● CPP-016 (Metadata Ingest and Management) ● CPP-017 (Disposal) ● CPP-026 (File Normalisation)		

Rationale(s)[1] and worst case(s)

Rationale	Impact of inaction or failure of the process
Quality assessment identifies vulnerabilities before they result in data loss, allowing the TDA to take preventive action rather than reactive measures. As digital preservation spans decades or centuries during which technological environments will change completely multiple times, the data quality assessment evaluates whether current *Objects* contain sufficient technical and contextual information to remain interpretable by future systems and users.	Uncontrolled file format obsolescence, hardware failure or bit corruption. Loss of content interpretability over time, authenticity and/or significant properties.
Data Quality Assessment helps the TDA to take informed preservation decisions regarding appraisal and re-appraisal based on quality metrics. Identification, automated extraction and correct interpretation of such metrics is fundamental to collection development.	No knowledge or no capacity to assess the quality of the *Object* could lead to appraisal of *Representations* of poor quality despite *Representations* of better quality being available.

2. Dependencies and relationships with other CPPs

Dependencies

CPP-ID	CPP-Title	Relationship description
CPP-007	Virus Scanning	Virus Scanning acts as a supplier since scanning for viruses is performed as a step in the overall Data Quality Assessment.
CPP-009	Metadata Extraction	Metadata extraction returns *Metadata* that are used to assess the *File* quality (e.g. for an audiovisual *File* quality assessment may rely on *Metadata* such as bit depth, sampling frequency, etc.)

[1] Term derived from PREMIS.

CPP-018	Community Watch	The signals from the community may affect the Data Quality Assessment. For example, the Data Quality Assessment performed during Ingest may result in extraction of quality properties that are required by the Designated Community.
CPP-020	Rights Management	Soft dependency (i.e. may require): Assessing the legal status and authenticity of *Objects* requires *Rights metadata*.
CPP-005	Identifier Management	Soft dependency (i.e. may require): Data Quality Assessment may include validating the PIDs and their linked resources.

Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Triggers	CPP-017	Disposal	The Data Quality Assessment tasks may discover intolerably low-quality issues in an *Object* and provide a trigger for disposal.
Required by	CPP-019	Metadata Extraction	The selection of an appropriate extractor tool depends on requirements as provided by Data Quality Assessment.
Required by	CPP-029	Ingest	Ingest uses the Quality Assessment report as produced by Data Quality Assessment to accept or reject the *Object*.
May be required by	CPP-029	Ingest	The TDA may have quality requirements as produced by Data Quality Assessment that may be checked during ingest.
Affinity with	CPP-013	Object Management Reporting	Object management reporting relates to re-evaluating quality dimensions.
Affinity with	CPP-022	Significant Properties Definition	As Data Quality Assessment identifies quality properties whose value will determine whether the *Objects* are ingested or not, these quality properties will likely be also considered significant by the TDA.

Affinity with	CPP-023	Risk Definition and Extraction	Both CPP-019 and CPP-023 are defining properties that the TDA should consider and interpret against the result of CPP-009 (Metadata extraction).
Affinity with	CPP-025	Enabling Access	*DIPs* should conform to the quality aspects as specified by the TDA.

3. Links to frameworks

Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS Link	Quality Assurance	R10 Quality Assurance
Nestor Seal Link	Quality Assurance	The question of quality assurance is mentioned in C22 Transformation of the submission information packages into archival information packages, C23 Archival information packages, C24 Interpretability of the archival information, C25 Transformation of archival information packages into dissemination information packages, C26 Dissemination information packages
ISO 16363 Link	Quality control	3.3.2.1 The repository shall have mechanisms for review, update, and ongoing development of its Preservation Policies as the repository grows and as technology and community practice evolve

Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS Link	Quality Assurance	4.2.2 General 4.2.3.3 Ingest Figure A-1: Composite of Functional Entities
PREMIS Link	/	/

4. Reference implementations

Example use case(s)

Access Quality Metrics for Net Art

Institutional Background	
Institution	Rhizome, USA
Hyperlink	https://doi.org/10.17605/OSF.IO/6RNK4
Description	
Trigger event	Rhizome’s ArtBase faces significant challenges in providing high-quality, reliable access to their collections. Over time, the software, hardware, and file formats used to create and view these works become obsolete, leading to a degraded user experience or rendering the art inaccessible.
Problem statement	The primary problem is the lack of a standardised method to help users of the ArtBase archive navigate the various versions and access methods of digital artworks. The archive holds multiple "variants" of each piece, which might include live versions from a web server, archived copies, or versions viewed through emulators. Each variant offers a different experience, and without a guide, users might unknowingly choose a version that is incomplete or partially non-functional. The paper notes that visitors need a way to make an informed choice between a version that is integrated into the modern internet landscape but potentially broken, and one that is more historically accurate but requires a special, emulated environment.
Proposed solution	The proposed solution is a system that calculates an "access quality score" for each variant of an artwork. This score is a single value, derived from a combination of *Technical metadata* and curatorial information, which indicates how complete and functional an artwork's performance is likely to be. The system uses a data model to define variants as a combination of archived *Files* ("artifacts") and the software environment ("machine") used to view them. The score is calculated by determining whether a machine's capabilities support the data formats within the artifact. This system aims to present a simple, three-level "stoplight" indicator (green, yellow, or red) that guides visitors to the best available version, manages their expectations for works with known issues, and ultimately improves the user experience of the ArtBase archive.

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/93608984/ Specifications TIB Pre-Ingest Analyzer (PIA): https://github.com/TIB-Digital-Preservation/pre-ingest-analyzer
	Non-commercial digital preservation service		
	Research infrastructure		
	Research performing organisation		
CSC – IT Center for Science Ltd., Finland	Non-commercial digital preservation service	English	https://digitalpreservation.fi/en/specifications
Archivematica	Digital preservation system	English	Manual assessment can be done using the Appraisal Tab: https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/appraisal/appraisal/ Some of the information needed for quality assessment (e.g. file format validation, characterisation, virus scanning) is produced during the Transfer process: https://www.archivematica.org/en/docs/archivematica-1.17/user-manual/transfer/transfer/#transfer-tab-microservices