International Standard

ISO/IEC 5259-3

First edition
2024-07

Artificial intelligence — Data quality for analytics and machine learning (ML) —

Part 3: Data quality management requirements and guidelines

Intelligence artificielle — Qualité des données pour les analyses de données et l'apprentissage automatique —

Partie 3: Exigences et lignes directrices pour la gestion de la qualité des données

Reference number
ISO/IEC 5259-3:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

A list of all parts in the ISO/IEC 5259 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

The quality of analytics and machine learning (ML) based products and services depends on the quality of data used to train ML models. Hence, data quality management is essential as it often helps to ensure the success of analytics and ML technology.

The adoption of a data quality management system facilitates managing the quality of products and services that employ analytics and ML technologies. This document defines vocabulary, requirements and guidelines for communication, alignment and agreement for managing data quality. The data quality management system provides transparency and auditability, either through self-assessment or third party assessment. It facilitates achieving relevant stakeholder satisfaction and managing quality, performance and self-declaration requirements. Specifically, this document defines requirements for a data quality management system with references to data quality measures that are relevant for the most commonly used analytics and ML technologies.

As data quality requirements vary with context and application domain, this document provides a generic set of requirements and recommendations relating to common data life cycle stages. A data life cycle is typically tightly integrated with the accompanying AI system life cycle and therefore has several dependencies. This document does not prescribe what AI system life cycle to use. Instead, it provides generic interfaces that allow users of this document the flexibility to interface with several life cycle models as long as the life cycle processes can be mapped.

ISO/IEC 5259-1 describes the data quality terminology and concepts used in this document.

ISO/IEC 5259-2[1)] describes the data quality model and data quality measures used in this document.

ISO/IEC 5259-4 describes the data quality process framework used in this document.

ISO/IEC 5259-5[2)] provides a data quality governance framework as guidance for governing bodies.

ISO/IEC TR 5259-6[3)] describes a visualization framework for data quality in analytics and ML.

1) Under preparation. Stage at the time of publication: ISO/IEC FDIS 5259-2:2024.

2) Under preparation. Stage at the time of publication: ISO/IEC DIS 5259-5:2023.

3) Under preparation. Stage at the time of publication: ISO/IEC CD TR 5259-6:2023.

Artificial intelligence — Data quality for analytics and machine learning (ML) —

Part 3:
Data quality management requirements and guidelines

1 Scope

This document specifies requirements and provides guidance for establishing, implementing, maintaining and continually improving the quality of data used in the areas of analytics and machine learning.

This document does not define a detailed process, methods or metrics. Rather it defines the requirements and guidance for a quality management process along with a reference process and methods that can be tailored to meet the requirements in this document.

The requirements and recommendations set out in this document are generic and are intended to be applicable to all organizations, regardless of type, size or nature.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 5259-1:2024, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples*

ISO/IEC 5259-2[4)], *Artificial Intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures*

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989, ISO/IEC 5259-1 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
data quality claim
statement to what degree data satisfy a data quality requirement

4) Under preparation. Stage at the time of publication: ISO/IEC FDIS 5259-2:2024.

3.2
data quality plan
specification of practices, processes and allocation of resources to achieve data quality objectives as the outcome of data quality planning

3.3
data quality planning
part of data quality management focused on setting data quality objectives and specifying necessary operational processes and related resources to achieve the quality objectives

[SOURCE: ISO 8000-2:2022, modified — example removed]

3.4
development interface agreement
DIA
agreement between customer and supplier in which the responsibilities for activities to be performed, evidence to be reviewed, or work products to be exchanged by each party related to the development of items or elements are specified

Note 1 to entry: While DIA applies to the development phase, supply agreement applies to production.

[SOURCE: ISO 26262-1:2018]

4 Symbols and abbreviated terms

DQMLC data quality management life cycle

5 Intended usage

This document may be used in one or more of the following modes:

— by an organization to establish and tailor a data quality management process for the use of data in analytics and ML, and continually improve processes;

— by an ML project to define, trace and evaluate data quality requirements;

— by a data user and data holder to establish a common understanding of data quality characteristics, and to ensure that agreed requirements have been met, facilitating an agreement for transacting data.

NOTE The organization can request assurances of confidentiality and proper use for supporting evidence.

6 Overall data quality management

6.1 Objective

The objective of a data quality management process is to establish appropriate (i.e. repeatable and auditable) processes to manage the quality of data and reliably meet a given set of requirements set by the organization.

6.2 General

Data quality impacts outcomes of analytics and ML algorithms. Data quality has an inherent constituent and a system-dependent constituent. Data can be suitable for one application but not suitable for another. This document helps to establish and maintain data quality for each analytics and ML application.

6.3 Requirements and recommendations

6.3.1 General

The following requirements and recommendations apply to the whole organization.

6.3.2 Data quality culture

The organization should sustain a data quality culture.

The organization shall:

a) have rules and processes to achieve quality (according to this document) taking into account the data quality model as applied to the applicable products and services;

b) define and implement data quality management processes, and perform related data quality activities;

c) integrate the data quality management processes and activities, to the extent appropriate, into other management processes and activities, such as general quality management and risk management;

d) document the performed activities;

e) provide resources sufficient to perform data quality management;

f) monitor, and to the extent necessary, review and improve the data quality management processes;

g) provide the required authority to involved personnel;

h) communicate data quality policies within the organization.

6.3.3 Management of data quality issues

The organization shall meet data quality requirements by:

a) having processes for communicating, analysing, evaluating, resolving and closing data quality issues;

b) documenting closed issues;

c) escalating or delegating issues that cannot be closed.

NOTE 1 Resolving and closing issues of data quality can include limiting or adjusting the scope of the ML project.

NOTE 2 A data quality issue can be closed by implementing a resolution or determining a resolution based on defined acceptance criteria.

6.3.4 Competence management

The organization shall manage competence by:

a) documenting required skills and tools to process the data;

b) ensuring that involved personnel have sufficient skills to perform their activities and duties;

c) maintaining records of persons and their proficiency on the required skills and tools;

d) keeping appropriate records of training and experience that substantiate claims of appropriate skills.

The organization can use external sources of competencies.

6.3.5 Resource management

The organization shall provide the resources required for data quality management, including:

a) software applications, training and support necessary to perform data quality management;

b) IT infrastructure or services necessary to perform data quality management (e.g. compute, storage, networking);

c) personnel with the skills required to perform data quality management.

6.3.6 Management system integration

The organization should integrate its data quality management activities into its existing management system, including its management systems for product or service quality, and for the development and use of AI systems. Implications from dual roles of stakeholders should be managed by the quality management system, including mitigation of any conflicts of interest.

NOTE 1 Stakeholder management can consider the potential of multiple roles for an individual. A user of analytics and ML based products or services can also be an owner or contributor of data.

NOTE 2 Organizations can use ISO/IEC 42001 to define a management system for the development or use of AI systems.

NOTE 3 Organizations can use ISO 9001 or other sector-specific quality management systems to define their quality management system.

6.3.7 Documentation

Documentation shall be intelligible to relevant stakeholders of the project in accordance with their role. Resources in a language that is not understood by a relevant stakeholder should be accompanied by a summary in a language that the stakeholder can understand.

The documentation shall be accessible to relevant stakeholders as appropriate and authorized. Access overhead should be minimized.

Documentation should include the context or references necessary to make it intelligible to future relevant stakeholders who are not part of the current project. This practice can enable these stakeholders to evaluate a dataset for potential reuse, partially or in total.

6.3.8 Data quality audit and assessment

The implemented processes shall be audited when appropriate, which shall be based on an evaluation of:

a) the data quality plan against organizational rules and processes;

b) arguments and justifications detailing how the requirements of the data quality model have been applied;

c) arguments detailing how the objectives of data quality plan have been achieved;

d) whether the data quality plan and all work products are complete, consistent and correct according to this document;

e) recommendations for improvement of data quality.

The data shall be assessed using a data quality assessment, which shall be based on an evaluation of whether the data achieve the objectives of this document, the current state-of-the-art in technology and the applicable engineering domain knowledge.

The data quality assessment plan shall be included in the specification stage. The data quality assessment shall be performed before data provisioning (see Figure 1, Stage 7: Data provisioning) or at an appropriate interval when using continuous learning or when using streaming data.

The data quality assessment may be performed on a subset of the data when it can be demonstrated that the quality of the subset is representative of the quality of the complete dataset.

6.3.9 Confirmation review and data quality measures

Data quality shall be confirmed by appropriate data quality measures in accordance with ISO/IEC 5259-2. A data quality review shall at least cover:

a) confirmation reviews of key work products. Every confirmation review:

   1) shall be finalized before data provisioning;

   2) should be based on whether the objectives of this document are achieved;

b) quality audits of the implemented processes;

c) quality assessment of the data.

All work products shall undergo confirmation reviews.

The personnel performing these reviews shall have access to the involved personnel, relevant information and required resources.

NOTE Confirmation reviews of key work products can be delegated, but the responsibility stays with the designated person.

6.3.10 Project-specific data quality management

The organization shall manage project-specific data by:

a) establishing a suitable project-specific data quality management process that meets all requirements of the specific ML project;

b) maintaining a list of relevant data quality claims. Where applicable, quantitative and qualitative benchmarks for data quality measures shall be documented;

c) adopting appropriate processes to identify and manage all data quality measures relevant for the project.

The project-specific data quality management process shall fulfil the requirements of Clause 12.

6.4 Work products

Work products of the data quality management process shall include:

a) organization-specific rules and processes for data quality (e.g. according to ISO/IEC 5259-4);

b) evidence of competence management;

c) evidence of a data quality management system;

d) identification of the used data quality measures;

e) documentation of applicable data quality measure benchmarks;

f) identified quality anomaly reports.

7 Life cycle-specific data quality management

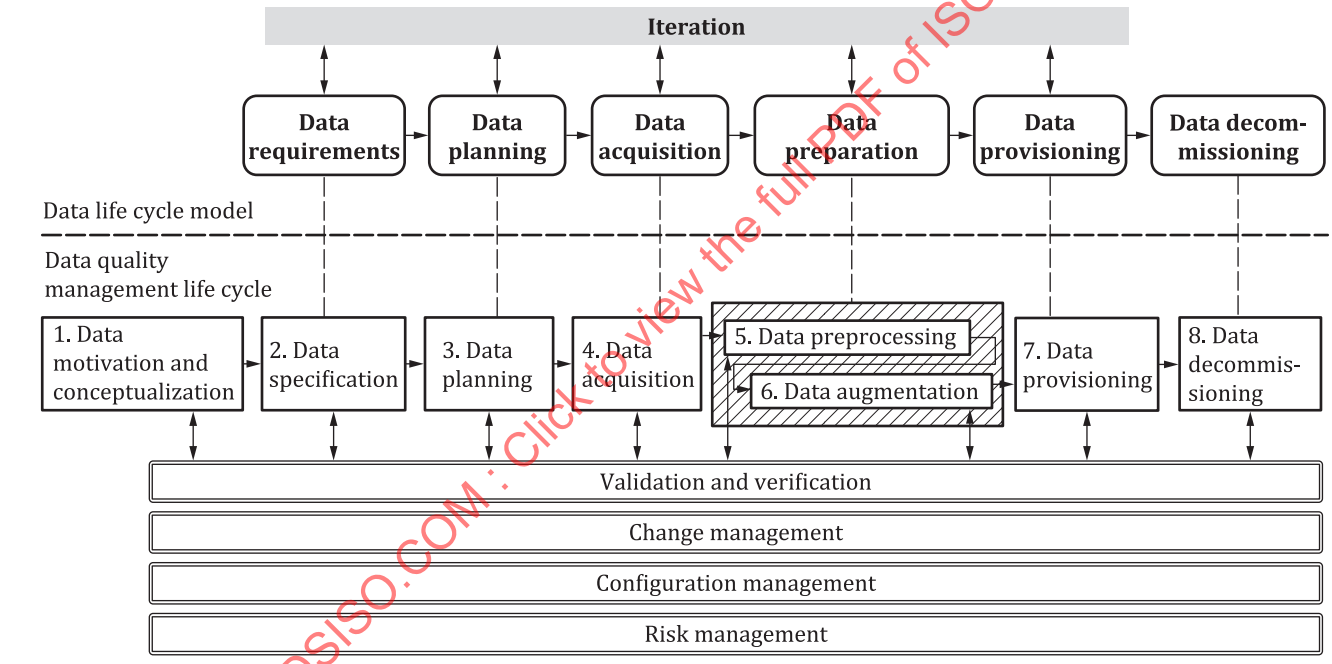
7.1 Objective

The objective of a data quality management life cycle (DQMLC) is to establish and maintain data quality throughout the data life cycle. An example of a data life cycle model is described in ISO/IEC 5259-1:2024, Figure 3.

7.2 General

7.2.1 Data quality management life cycle

Data quality shall be managed in all stages of the data life cycle. The data quality management life cycle (DQMLC) model shown in Figure 1 provides guidance towards meeting the quality requirements of data for use in an analytics and ML context. It derives discrete stages that are relevant for the management of data quality and facilitates grouping and organizing of requirements and guidelines to consider for managing data quality. The DQMLC model is not prescriptive of the temporal ordering of stages. The various stages of the DQMLC are described in 7.2.2.



Key

- Primary development pathway
- Feedback pathway
- Instantiation of DLC model for 5259 series (see ISO/IEC 5259-1:2024, Figure 3)
- Data quality management life cycle stage
- Horizontal processes of the data quality management life cycle
- Indicating iteration within the life cycle
- Grouping of data quality management life cycle stages to clarify mapping to DCL model

Figure 1 — Data quality management life cycle

Issues in data quality can originate at any stage throughout the data life cycle. Data quality management shall establish and maintain data quality processes at the beginning of the data life cycle. If the organization delegates the responsibility for a process, that delegation shall be documented and be traceable.

NOTE It is generally more difficult to detect and correct data quality issues after the fact than to manage data quality risks when they first occur. For example, errors introduced while collecting data are more easily avoided by appropriate quality management rather than attempting to detect and correct errors later in the data life cycle.

7.2.2 Data quality management life cycle stages

7.2.2.1 Data motivation and conceptualization

Data quality management begins with the motivation and conceptualization stage. Potential issues with data quality should be identified and managed when the first need for data for an analytics and ML context becomes apparent. Among others, the motivation and intended use for the data are validated and verified to manage quality characteristics such as conformity and relevance.

7.2.2.2 Data specification

In the data specification stage, data quality management imposes requirements on the creation of the data specification, including data formats, statistical properties and divisibility. Data quality management guidelines facilitate identification of erroneous, incomplete, or conflicting requirements and plans. For example, given an analytics and ML context, quality management confirms that the data match the requirements of that context.

7.2.2.3 Data planning

In the data planning stage, a plan is formulated to meet the data specification. This includes planning of specific tasks and resources to acquire and process data throughout the data life cycle, as well as evaluation methods and acceptance criteria.

7.2.2.4 Data acquisition

In terms of data quality management, data acquisition can refer to collecting data, generating data, or acquiring and combining existing datasets. The choice of method of data collection has implications on data quality management. Regardless of the method of obtaining data, the result of this stage is considered the raw input for subsequent stages of the data life cycle, even if an already managed dataset is imported.

7.2.2.5 Data preprocessing

The data preprocessing stage organizes all data quality management requirements and guidelines that are related to processing the raw input data obtained during the data acquisition stage. Activities in this stage tend to remove data items, for example by filtering and cleaning raw data.

7.2.2.6 Data augmentation

Data augmentation takes the input from the data preprocessing stage and organizes it according to all quality management requirements and guidelines related to adding to the dataset. This includes associating metadata, adding data labels or other transformations.

7.2.2.7 Data provisioning

The goal of data provisioning is to provide the dataset where it is needed. Data quality management requirements in this stage include, among others, maintaining integrity of datasets during delivery, access control, documentation and verification that context of use matches the dataset's intended use.

7.2.2.8 Data decommissioning

Data quality management closes with data deletion or data transfer in the data decommissioning stage. In each of these mutually exclusive cases, the responsibility for the data is released. The appropriate handling of data released is a major factor of trust in the organization and can contribute to data misuse prevention.

7.2.3 Project-independent tailoring of the data quality management life cycle

The organization should tailor the life cycle if it:

a) joins or splits the stages;

b) performs required processes in different or added stages (see ISO/IEC 5259-4);

c) iterates stages;

d) performs activities concurrently if they do not depend on each other;

e) omits stages that do not apply, given a rationale;

f) includes data quality management provisions for validation and verification, change management and configuration management as shown in Figure 1.

7.2.4 Horizontal aspects of the data quality management life cycle

7.2.4.1 General

Figure 1 shows four processes that cover all stages of the data quality management life cycle including:

a) validation and verification;

b) change management;

c) configuration management;

d) risk management.

These processes shall also be carried out in the case of project-independent tailoring.

7.2.4.2 Validation and verification

Validation and verification of the data quality should be done at all relevant stages throughout the data quality management life cycle.

7.2.4.3 Change management

Change management ensures that any alteration to the data or processes does not adversely affect data quality.

7.2.4.4 Configuration management

Configuration management helps to ensure that data and processes are uniquely referenceable, and that data quality is not adversely affected by any configuration management changes.

7.2.4.5 Risk management

Risk management should be done at all relevant steps throughout the data quality management life cycle.

7.3 Requirements and recommendations

7.3.1 Data motivation and conceptualization

7.3.1.1 General

The organization shall:

a) document motivation and origin of the data need;

b) specify scope, purpose and intended use of data;

c) document data quality requirements.

NOTE For more information on societal and ethical concerns see ISO/IEC TR 24368.

7.3.1.2 Stakeholder analysis

The organization shall:

a) perform a stakeholder analysis to identify all relevant stakeholders of the data and implications for data quality management;

b) identify potentially conflicting data quality requirements.

NOTE A description of potential AI stakeholder roles and sub-roles is provided by ISO/IEC 22989:2022, 5.19.

7.3.1.3 Feasibility analysis

The organization should maintain a feasibility analysis based on its available skills and resources that assesses the organization's ability to meet its data quality goals. A feasibility analysis should be updated throughout the data life cycle.

NOTE Prior ML projects and openly available datasets that have achieved target quality goals are examples of evidence supporting feasibility.

7.3.2 Data specification

7.3.2.1 General

The organization shall specify data requirements in a data specification and validate that these requirements are consistent and complete for the intended use.

The data specification shall include which aspects are minimally required, and which are optional based on the intended use by the AI system.

The specification should cover:

a) a description of the nature of the data;

b) the intended goal of the analytics and ML project using the data;

c) legal requirements;

d) safety and security considerations;

e) potential of unwanted bias;

f) privacy considerations;

g) domain and project-specific data requirements;

h) data quality model and required quality level to achieve intended use of data.

NOTE 1 Data specification can be based on the ISO/IEC TR 24027 description of treatment of unwanted bias throughout an AI system life cycle.

NOTE 2 Data specification includes the various interfaces with an ML project (e.g. data for training, data for testing or ongoing data management).

NOTE 3 Some ML algorithms require specific statistical properties of the training data.

NOTE 4 Privacy considerations can include the principle of data minimization (also-known-as data avoidance). Data minimization means that only the minimum data necessary for the function of the system are collected or used.

7.3.2.2 Data format

The organization shall determine which information is to be included in the data format information. Such information can include, but is not limited to:

a) encoding;

b) frequency (time);

c) resolution (space);

d) syntax;

e) semantics;

f) structure of associated metadata;

g) expected value ranges;

h) optional and mandatory elements;

i) optionally allowable property formats and valid alternatives (e.g. image sizes);

j) links and cross-references.

NOTE Semantics refers to interpretation and use of the data, which includes allowable operations on the data.

7.3.2.3 Statistical properties and divisibility

The data specification shall include specification of valid divisions into subsets (e.g. training, validation and test data). The subsets should not intersect except when used in conjunction with scientifically appropriate methods such as bagging, bootstrapping and k-fold validation.

The organization shall determine and maintain for each of these subsets:

a) relevant statistical properties;

b) representativeness;

c) appropriate size.

7.3.2.4 Supporting resources and tools

The data specification should include all relevant requirements on data handling and at least cover:

a) supporting tools specifications;

b) minimal hardware specification to process data and operate tools;

c) necessary resolutions for visualization;

d) storage requirements;

e) network requirements;

f) accessibility requirements.

7.3.3 Data planning

7.3.3.1 General

The organization shall establish a data quality plan that covers processes, activities and resources to meet the data specification throughout the data life cycle.

The data quality plan shall include:

a) data quality goals in alignment with the data specification;

b) committed resource allocation;

c) provisions to meet legal;

d) processes to manage and update the data quality plan when needed;

e) provisions to ensure that processes are traceable and reproducible.

7.3.3.2 Data life cycle specific plans

The data quality plan shall include process specifications, including description, related activities, intended outcomes with an assignment to a responsible person for each process, covering the following data life cycle stages:

a) data acquisition plan;

b) data preprocessing plan;

c) data augmentation plan;

d) data provisioning plan;

e) data decommissioning plan.

The data quality plan corresponding to the data life cycle stages shall meet the requirements as stated in 7.3.4 through 7.3.8.

7.3.4 Data acquisition

7.3.4.1 General

The organization shall specify data acquisition processes to meet the data specification and include them in the data quality plan.

Appropriate aspects of the data specification (see 7.3.2) should be completed prior to data preprocessing. Risks that arise from incompleteness of the data shall be documented and mitigated. The organization shall monitor data acquisition processes and document deviations from expected outputs or deviations from planned processes.

The organization acquiring the data shall take responsibility for the data, including their quality management.

NOTE Data acquisition processes can be applicable independent of data source.

7.3.4.2 Data source

When establishing the method of data acquisition, the organization should consider whether:

a) required data already exist and directly are available for reuse;

b) existing data can be transformed to meet requirements;

c) data can be purchased or licensed;

d) new data need to be collected;

e) data can be generated (e.g. through simulation or other computational means).

The organization can delegate data acquisition, but the organization remains responsible for meeting the organization's data quality requirements. Documented conditions of the reuse or license of a dataset shall be taken into account.

NOTE Available open data are a viable option for a growing number of applications and promote data reuse. Existing datasets already owned by the organization can also be considered for reuse.

7.3.4.3 Data collection

In case the organization needs to collect new data, the organization shall provide:

a) reasons for collection of new data;

b) method of data collection (e.g. through sensors, manual data entry, derivative value, simulation, synthetic data);

c) data collection specification, including:

   1) relevant configuration and parameter settings of data collection methods;

   2) operational conditions;

   3) error detection and mitigations;

   4) required skills and resources;

   5) if applicable, specification and mounting location of sensors;

d) required resources and skills for data collection and handling;

e) methods for anonymization or pseudonymization if applicable;

f) documentation on data measurement scales (e.g. nominal, ordinal, interval and ratio) and measurement units;

g) expected deviations from target data quality.

NOTE The data collection method can introduce systematic errors to the collected data, especially if the data collection setup differs from the intended setup in use. One possible mitigation is the use of dissimilar methods of collection to provide redundancies.

7.3.4.4 Data handling

The organization shall establish appropriate processes and tools to manage, visualize and review the acquired data. Acquired data shall meet the data specification.

The organization shall specify processes to manage acquired data including:

a) quality review;

b) access control;

c) traceability of data origin and modifications;

d) version control;

e) data storage, location and backup;

f) dataset operations such as update, merge, sort and slice;

g) processes to detect, minimize and mitigate data corruption;

h) assignment of a responsible person for the data.

7.3.5 Data preprocessing

7.3.5.1 General

The organization shall specify data preprocessing processes to meet the data specification and include them in the data quality plan.

7.3.5.2 Data cleaning

The organization shall at least specify which data quality measures have been applied for the following:

a) detection and mitigation of missing data;

b) detection and mitigation of data duplication;

c) detection and mitigation of outliers and other issues;

d) detection and mitigation of bias, drift and scaling;

e) detection and handling of non-standard values;

f) detection and removal of not needed data;

g) data transformations, including data normalization;

h) methods of verifying data cleaning results;

i) if applicable, data de-identification.

NOTE Some data transformations are not reversible.

7.3.6 Data augmentation

7.3.6.1 General

The organization shall specify data augmentation processes to meet the data specification and include them in the data quality plan.

7.3.6.2 Data labelling and annotation

The organization should specify data labelling, including:

a) data labelling specification;

b) required skills and resources;

c) selection of data for labelling;

d) monitoring and quality management of labelling processes;

e) potential physical and psychological impact on data labeller, including mitigation strategies.

7.3.6.3 Computed augmentation

If computed augmentation is used, the organization should specify:

a) tools and methods used for computed augmentation;

b) selected features;

c) created logs and metadata.

7.3.7 Data provisioning

7.3.7.1 General

The organization shall:

a) specify data provisioning processes to meet the data specification and include them in the data quality plan;

b) provide auditable evidence that the provisioned data and metadata meet all specified requirements;

c) implement appropriate processes to ensure that items and versions described in 7.3.7.2 are provided;

d) ensure the provisioning does not alter the data;

e) implement appropriate authentication and data access processes;

f) version control all provisioned items;

g) package and deliver the data including documentation and other associated items so that the user of the data is able to:

   1) determine if the data meet usage requirements;

   2) use the data according to their intended use.

NOTE For example, data representation that stores a large number of decimals can be misleading by suggesting a higher than achieved accuracy and hence can cause unintended usage or assumptions for use.

7.3.7.2 Data provision items

The following items should be included in the data provision description:

a) the data including any splits, if appropriate, to produce training, testing and validation data;

b) documentation (see 7.3.7.3);

c) data sample representative of data syntax and format;

d) documentation of data specification, other associated items (see 7.3.7.1,g), data quality model, data quality measures and the results of the data quality assessment.

7.3.7.3 Documentation

The following documentation shall be provisioned with the data:

a) scope of use;

b) instructions for use;

c) statistical properties of the data;

d) all data item and metadata descriptions;

e) relevant stakeholders of the data;

f) roles and duties for the recipient, including legal requirements and contractual obligations;

g) requirements for use environments; typically covering testing, development and system operation environments (see Stage 8 in ISO/IEC 5259-1:2024, Figure 3);

h) factors that can impact quality measures negatively;

i) available acceptance test methods and their expected results;

j) communication strategy with the data user, including how updates occur and how users are notified.

7.3.7.4 Tracking and improvement

The organization shall evaluate if there is a requirement to track usage and results of usage of the data. In case such requirements exist or the organization decides to implement tracking, the organization shall implement appropriate processes to meet tracking and improvement requirements.

In case data usage tracking is implemented, the organization shall evaluate the need to:

a) update documented scope and intended use of the data;

b) improve the data;

c) update data quality management system.

The organization shall establish a channel to collect voluntary feedback from relevant stakeholders of the data and establish processes to evaluate the feedback for continual improvement of the data and the data quality management system.

7.3.7.5 Data optimization

The organization shall document applied data optimizations:

a) storage space optimizations;

b) access optimizations;

c) networking resource requirement optimizations;

d) data handling optimizations (e.g. buffering, caching).

7.3.7.6 Support commitment

The level of support provided to relevant stakeholders shall be sufficient to achieve the data quality requirements in the context of the intended use of the data. The following items shall be documented:

a) required skills or trainings to use data;

b) documentation of support provided, including scope, period and availability;

c) maintenance plan and commitment to continuously improve data;

d) available communication channel with relevant data stakeholders.

7.3.8 Data decommissioning

7.3.8.1 General

The organization shall choose an appropriate method to decommission the data either through data deletion or through data transfer.

The organization shall specify the data decommissioning processes to meet all relevant requirements (see 7.3.8.2–7.3.8.4) and include them in the data quality plan.

7.3.8.2 Data transfer

In case of data transfer, the organization shall identify a recipient of the data that is qualified to take over the data and all associated legal requirements.

The organization shall document that data transfer is not violating any obligations.

The organization shall document an expressed agreement from the recipient to take over the data and all associated obligations.

The organization shall generate and independently validate a data transfer report.

7.3.8.3 Data deletion

The organization shall document that data deletion is not violating obligations, including, but not limited to, legal requirements and data retention obligations.

The organization shall ensure that there is no data user that still requires access to the data. Data users can include systems that require access to the data for maintenance and re-training.

In case of deletion, the organization shall delete the data from all storage locations and document data deletion methods and their security (e.g. overwriting disk with random data).

The organization shall review if there is potential to restore the data either partially or fully from ML models that have been trained with the data. The organization shall consider such cases in its obligations and data decommissioning plan.

The organization shall generate and independently validate a data deletion report.

The organization should review if the data are of cultural, historical or societal significance. In case the data are of significance, the organization should document efforts to transfer the data to a recipient that can maintain all data obligations and their significance.

The organization should review if it is possible and there is value to donate the data into the public domain.

NOTE 1 Storage locations of data include backups that can be automatically generated by the IT infrastructure.

NOTE 2 It is good practice for the organization to provide an argument why a data transfer harms its interest to support the organization's choice of data deletion.

NOTE 3 Donating data to the public domain potentially benefits research and education and indirectly the organization itself.

7.3.8.4 Partial data deletion on request

The organization shall have appropriate processes in place to handle partial data deletion requests, if applicable.

In case parts of the data are deleted, the organization shall evaluate if the data that remain still meet their specifications. In case the data do not meet their specifications, the organization shall notify relevant stakeholders of the risks and issues in quality and provide appropriate mitigation if possible. In case no

appropriate mitigation can be implemented, the organization shall decommission all of the data, which can include transfer of the remaining data.

In case of partial data deletion, the organization shall delete the partial data from all storage locations.

The organization shall generate and independently validate a partial data deletion report.

NOTE Data regulations such as privacy protection laws can compel the organization to delete parts of a dataset.

7.4 Work products

7.4.1 Work products of data motivation and conceptualization stage

Work products of the data motivation and conceptualization stage include:

a) documentation of intended use of data;

b) feasibility review;

c) stakeholder analysis;

d) evidence that verification and validation of data motivation and conceptualization stage have been completed successfully.

7.4.2 Work products of data specification stage

Work products of the data specification stage include:

a) data format and use specification;

b) specification of supporting tools and associated requirements;

c) evidence that verification and validation of data specification stage have been completed successfully.

7.4.3 Work products of data planning stage

Work products of the data planning stage include:

a) data quality plan;

b) evidence that verification and validation of data planning stage have been completed successfully.

7.4.4 Work products of data acquisition stage

Work products of the data acquisition stage include:

a) collected data and associated documentation;

b) specification of methods of data acquisition and related configurations;

c) infrastructure to manage data;

d) data quality review;

e) evidence that verification and validation of data acquisition stage have been completed successfully.

7.4.5 Work products of data preprocessing stage

Work products of the data preprocessing stage include:

a) cleaned data;

b) data quality review;

c) descriptive statistics describing properties of the data;

d) evidence that verification and validation of data preprocessing stage have been completed successfully.

7.4.6 Work products of data augmentation stage

Work products of the data augmentation stage include:

a) augmented data, including metadata and labels;

b) feature descriptions;

c) specification of applied augmentation tools and methods;

d) evidence that verification and validation of data augmentation stage have been completed successfully.

7.4.7 Work products of data provisioning stage

Work products of the data provisioning stage include:

a) provisioned items, including data, data sample and documentation;

b) evidence that verification and validation of data provisioning stage have been completed successfully.

7.4.8 Work products of data decommissioning stage

Work products of the data decommissioning stage include:

a) data decommissioning report;

b) evidence that verification and validation of data decommissioning stage have been completed successfully.

8 Horizontal processes

8.1 Objective

The objective of data quality management horizontal processes is to consolidate activities that are applicable to each stage of the data quality management life cycle.

8.2 General

The organization shall apply and document horizontal processes with process-specific inputs and outputs.

8.3 Requirements and recommendations

8.3.1 Verification and validation

8.3.1.1 General

Verification and validation processes monitor and assess data quality throughout all stages of the data quality management life cycle. Verification establishes objective evidence that data quality requirements have been met. Validation establishes that the data meet intended objectives.

The constraints and limitations of verification and validation processes shall be documented and the potential impact on data quality shall be evaluated.

8.3.1.2 Data life cycle quality gates

Each data quality life cycle stage shall specify objectively assessable quality targets of its work products. Verification and validation shall confirm that all requirements in each stage of the data quality life cycle have been met. A data life cycle stage shall only be entered with verified and validated input. The results of verification and validation shall be documented.

8.3.1.3 Improvement

After successful passing verification and validation of a work product, available feedback from the work products in its intended use shall be documented and evaluated. A failure of a work product to meet intended use specification can result in an update of verification and validation processes.

The results of verification and validation shall be used for regular updates and improvements of data quality management processes.

8.3.2 Configuration management

Configuration management shall be planned and maintained throughout the entire data quality management life cycle.

Work products defined by the data quality plan shall be based on the configuration management strategy, which defines requirements and purposes for uniquely identified and reproducible items.

Configuration management shall establish applicable data quality requirements for each valid configuration.

Configuration management shall establish that all valid configurations are meeting all applicable data quality requirements.

8.3.3 Change management

8.3.3.1 General

Change of data quality requirements and processes is expected to occur frequently during development of AI systems and is supported in the data quality management life cycle using iterations. Data and associated metadata are a frequent target for changes, mostly during the specification and implementation stages, and to a lesser extent as a result during the validation and verification processes. Changes to the data can also occur by revisiting data acquisition as a result of activities in data provisioning.

8.3.3.2 Planning change management

Change management shall:

a) be planned and initiated before any changes;

b) include a plan that identifies items subject to change and defines a change schedule;

c) define a process that includes:

   1) change request specification (see 8.3.3.3);

   2) change request analysis (see 8.3.3.4);

   3) change request evaluation (see 8.3.3.5);

   4) documentation (see 8.3.3.6).

8.3.3.3 Change request specification

Change requests shall have:

a) a unique identifier;

b) a date;

c) reason, description and configuration of the requested change;

d) decisions and a rationale for the requested change.

8.3.3.4 Change request analysis

The impact of the change request shall be analysed, considering:

a) its type (error, adaptation, addition to data or labels, etc.);

b) affected items (data, labels, requirements, etc.);

c) affected parties, including responsibilities;

d) effect on quality;

e) effect on pre-existing data or labels;

f) the schedule.

8.3.3.5 Change request evaluation

The change request and its impact analysis shall be evaluated to decide on:

a) its status (accept, reject, etc.) by an authorized person;

b) a personnel to carry out the changes;

c) its timeline.

8.3.3.6 Implementing and documenting change

Changes shall be implemented and verified as planned.

Changes to data, its quality, usage or verification shall trigger an update to affected assessments and reviews.

Documentation for changes shall include:

a) list of changed items, including configuration and versions;

b) details of the performed changes;

c) date for changes to take effect.

8.3.4 Risk management

8.3.4.1 General

The objective of risk management is to ensure that all risks to data quality are accounted for, assessed and, if necessary, mitigated.

8.3.4.2 Requirements and recommendations

The organization shall conduct a risk assessment and maintain documentation with respect to risks associated to data quality. The risk assessment should consider reasonably expectable uses out of the intended use of data and their risk consequences. The organization shall manage and maintain appropriate processes to minimize identified risks.

Risks of the intended use of data shall be specified, justified and documented as part of actions taken by the organization to assess and treat their risk. The selection of relevant data quality measures shall be updated and managed by the data quality management system based on the associated risk. The organization shall communicate risks associated with the data to relevant stakeholders.

NOTE To define risk management processes appropriate for AI, organizations can consider ISO/IEC 23894.

8.4 Work products

8.4.1 Work products of verification and validation

Work products of verification and validation include:

a) documentation of verification and validation results in all life cycle stages;

b) documentation of constraints and limitations of verification and validation;

c) improvement suggestions.

8.4.2 Work products of configuration management

Work products of configuration management include:

a) documented configuration management plan;

b) documented configuration strategy;

c) documentation of valid configurations and associated quality requirements.

8.4.3 Work products of change management

Work products of change management include:

a) change management plan;

b) change requests;

c) change impact analysis;

d) change reports.

8.4.4 Work products for risk management

Work products of a risk assessment include:

a) documentation on data quality risks;

b) risk treatment plan.

9 Management of data quality in supply chains

9.1 Objective

The objective of supply chain management requirements is to ensure that any selected suppliers provide data that meet the requirements of the organization.

9.2 Requirements and recommendations

Supplier selection should consider supplier capabilities to produce quality data according to Clauses 6–8 of this document.

Requirements to the request for quotation include:

a) The request for quotation to the supplier should contain:
   1) requirements to comply with this document;
   2) the relevant specifications for the data;
   3) relevant quality requirements.

b) User and supplier should specify a development interface agreement with:
   1) appointment of user, supplier and quality managers;
   2) all activities in the data quality management life cycle to be performed by each party;
   3) shared information and work products;
   4) the responsibilities assigned to each party for each activity;
   5) the qualitative and quantitative benchmark requirements (see 6.3.10 item b));
   6) the interface-related processes, methods and tools;
   7) the quality assessment activities;
   8) the planning of the supplier's quality assessment report;
   9) the agreement that allows a user assigned auditor to access all required resources for the purpose of the quality audits;
   10) the responsibilities and activities during operation and decommissioning (see Stages 8–10 in ISO/IEC 5259-1:2024, Figure 3);
   11) requirements to communicate information and issues, which have an impact to data quality or the risk of violating the development interface agreement;
   12) data lifetime requirements for data availability and use.

9.3 Work products

Work products of data quality management in supply chains include:

a) supplier selection report;

b) development interface agreement;

c) quality assessment report.