International Standard

ISO/IEC 5259-1

First edition
2024-07

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

Part 1:
Overview, terminology, and examples

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

Partie 1: Vue d'ensemble, terminologie et exemples

Reference number
ISO/IEC 5259-1:2024(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

Data are the raw material for analytics and machine learning (ML) and data quality is a critical aspect for related analytics and ML projects and systems. The aim of the ISO/IEC 5259 series is to provide tools and methods to assess and improve the quality of data used for analytics and ML.

Other parts of the ISO/IEC 5259 series include:

- ISO/IEC 5259-2[1] provides a data quality model, data quality measures and guidance on reporting data quality in the context of analytics and ML. ISO/IEC 5259-2 builds on the ISO 8000 series, ISO/IEC 25012 and ISO/IEC 25024.

  The aim of ISO/IEC 5259-2 is to enable organizations to achieve their data quality objectives and is applicable to all types of organizations.

- ISO/IEC 5259-3 specifies requirements and provides guidance for establishing, implementing, maintaining and continually improving the quality for data used in the areas of analytics and ML.

  ISO/IEC 5259-3 does not define detailed processes, methods or measurement. 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 ISO/IEC 5259-3.

  The requirements and recommendations set out in ISO/IEC 5259-3 are generic and are intended to be applicable to all organizations, regardless of type, size or nature.

- ISO/IEC 5259-4 provides general common organizational approaches, regardless of type, size or nature of the applying organization, to ensure data quality for training and evaluation in analytics and ML. It includes guidelines on the data quality process for:

  - supervised ML with regard to the labelling of data used for training ML systems, including common organizational approaches for training data labelling;
  - unsupervised ML;
  - semi-supervised ML;
  - reinforcement learning;
  - analytics.

  ISO/IEC 5259-4 is applicable to training and evaluation data that come from different sources, including data acquisition and data composition, data pre-processing, data labelling, evaluation and data use. ISO/IEC 5259-4 does not define specific services, platforms or tools.

- ISO/IEC 5259-5[2] provides a data quality governance framework for analytics and machine learning to enable the governing bodies of organization to direct and oversee the implementation and operation of data quality measures, management, and related processes with adequate controls throughout the DLC model according to ISO/IEC 5259-1.

- ISO/IEC TR 5259-6[3] describes a visualization framework for data quality in analytics and ML. The aim is to enable stakeholders using visualization methods to access the results of data quality measures. This visualization framework supports data quality goals.

[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 1: Overview, terminology, and examples

1 Scope

This document provides the means for understanding and associating the individual documents of the ISO/IEC 5259 series and is the foundation for conceptual understanding of data quality for analytics and machine learning. It also discusses associated technologies and examples (e.g. use cases and usage scenarios).

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 22989, *Information technology — Artificial intelligence — Concepts and terminology*

ISO/IEC 23053, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989 and ISO/IEC 23053 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 life cycle
life cycle of data
stages in the process of data usage from idea conception to its discontinuation

3.2
data originator
party that created the data and that can have rights

Note 1 to entry: A data originator can be an individual person.

Note 2 to entry: The data originator can be distinct from the natural or legal person(s) mentioned in, described by, or implicitly or explicitly associated with the data. For example, PII can be collected by a data originator that identifies other individuals. Those data subjects (PII Principals) can also have rights, in relation to the data set.

Note 3 to entry: Rights can include the right to publicity, right to display name, right to identity, right to prohibit data use in a way that offends honourable mention.

[SOURCE: ISO/IEC 23751:2022, 3.2]

3.3
data holder
party that has legal control to authorize data processing of the data by other parties

Note 1 to entry: A *data originator* (3.2) can be a data holder.

[SOURCE: ISO/IEC 23751:2022, 3.4]

3.4
data user
party that is authorized to perform processing of data under the legal control of a *data holder* (3.3)

[SOURCE: ISO/IEC 23751:2022, 3.5]

3.5
data quality
characteristic of data that the data meet the organization's data requirements for a specified context

3.6
data quality characteristic
category of data quality *attributes* (3.13) that has a bearing on *data quality* (3.5)

[SOURCE: ISO/IEC 25012:2008, 4.4, modified — Definition revised.]

3.7
data quality model
defined set of characteristics which provides a framework for specifying data *quality requirements* (3.9) and evaluating *data quality* (3.5)

[SOURCE: ISO/IEC 25012:2008, 4.6]

3.8
data quality measure
variable to which a value is assigned as the result of *measurement* (3.10) of a *data quality characteristic* (3.6)

[SOURCE: ISO/IEC 25012:2008, 4.5, modified — Note to entry removed.]

3.9
quality requirement
requirement for quality properties or *attributes* (3.13) of an information and communications technology (ICT) product, data or service that satisfy needs which ensue from the purpose for which that ICT product, data or service is to be used

[SOURCE: ISO/IEC 25030:2019, 3.15, modified — Note to entry removed.]

3.10
measurement
set of operations having the object of determining a value of a measure

[SOURCE: ISO/IEC 25024:2015, 4.27]

3.11
measurement scale
quantity-value scale
ordered set of quantity values of quantities of a given kind of quantity used in ranking, according to magnitude, quantities of that kind

EXAMPLE 1

Celsius temperature scale.

EXAMPLE 2

Time scale.

EXAMPLE 3

Rockwell C hardness scale.

[SOURCE: ISO/IEC Guide 99: 2007, 1.28, modified — Preferred term swapped with admitted term.]

3.12
analytics
data analytics
composite concept consisting of data acquisition, data collection, data validation, data processing, including data quantification, data visualization, data documentation and data interpretation

Note 1 to entry: Analytics is used to understand objects or events represented by data, to make predictions for a given situation and to recommend steps to achieve objectives. The insights obtained from analytics are used for various purposes such as decision-making, research, sustainable development, design and planning.

[SOURCE: ISO/IEC 20546:2019, 3.1.6, modified — The term "analytics" added as a preferred term, definition and note to entry revised.]

3.13
attribute
property or characteristic of an object that can be distinguished quantitatively or qualitatively by human or automated means

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.2, modified — Definition revised.]

3.14
feature
<machine learning> measurable property of an object or event with respect to a set of characteristics

Note 1 to entry: Features play a role in training and prediction.

Note 2 to entry: Features provide a machine-readable way to describe the relevant objects. As the algorithm will not go back to the objects or events themselves, feature representations are designed to contain all useful information.

[SOURCE: ISO/IEC 23053: 2022, 3.3.3]

3.15
data quality management
coordinated activities to direct and control an organization with regard to *data quality* (3.5)

[SOURCE: ISO 8000-2:2020, 3.8.2]

3.16
data governance
governance of data
system by which the current and future use of data is governed

3.17
data provenance
provenance
information on the place and time of origin, derivation or generation of a dataset, proof of authenticity of the dataset, or a record of past and present ownership of the dataset

[SOURCE: ISO/IEC 11179-33:2023, 3.11, modified — The term "data provenance" added as a preferred term, definition revised.]

3.18
visualization
scientific visualization
<computer graphics> use of computer graphics and image processing to present models or characteristics of processes or objects for supporting human understanding

EXAMPLE A display image created by combining magnetic resonance scans of a tumour; volumetric top and side views of a lake showing temperature data; a two-dimensional model of electrical waves in the heart.

[SOURCE: ISO/IEC 2382:2015, 2125942, modified — Preferred term swapped with admitted term, note to entry removed]

3.19
machine learning project
ML project
project that utilizes *analytics* (3.12) and machine learning and is responsible for the associated data throughout the data's entire life cycle

3.20
data architecture
description of the structure and interaction of the enterprise's major types and sources of data, logical data assets, physical data assets and data management resources

Note 1 to entry: Logical data entities can be tied to applications, repositories and services and may be structured according to implementation considerations.

Note 2 to entry: The concept of "data" is intentionally not defined here, as it is part of the data architecture definition for each application scenario. It is according to the specific requirements of that scenario.

[SOURCE: ISO TR 21965:2019, 3.2.6]

3.21
data item
smallest identifiable unit of data within a certain context for which the definition, identification, permissible values and other information is specified by means of a set of properties

Note 1 to entry: "Field" is considered a synonym of data item.

Note 2 to entry: Data item is a physical object "container" of data values.

[SOURCE: ISO/IEC 25024:2015, 4.9].

3.22
data record
set of related *data items* (3.21) treated as a unit

[SOURCE: ISO/IEC 25024:2015, 4.15]

3.23
metadata
data that define and describe other data

Note 1 to entry: In the context of *analytics* (3.12) and machine learning, metadata provides information on *data items* (3.21) or data *records* (3.22) such as their properties, structure, type, context, intended use, ownership, access and volatility.

[SOURCE: ISO/IEC 11179-1:2023, 3.2.26, modified — Note to entry added.]

4 Symbols and abbreviated terms

AI	artificial intelligence
DL	deep learning
DLC	data life cycle
DQ	data quality
ETL	extract, transform and load
ML	machine learning
PII	personal identifiable information

5 Data quality concepts for analytics and machine learning

5.1 Data quality considerations for analytics and machine learning

5.1.1 General

Existing data quality standards, such as the ISO 8000 series, were developed from the perspectives of data production and management. This is because data producers (or data collectors) were traditionally the largest consumers of data. Since most of the data were used for a predetermined purpose and associated data quality standards focused on only the characteristics necessary for the defined purpose, data produced in that manner can require additional processing for use in other contexts.

In the field of data analysis and ML, data users are generally not producing data. They search, collect and process data they believe are necessary and suitable for their analytics and ML project. In this case, data quality has an impact on the quality of the analysis results and the performance of the ML model. No matter how good the data analysis or ML model is, the results can be unreliable when using data that does not meet requirements. Even when data meets requirements for a particular application or context, it does not necessarily meet requirements for other applications or contexts. Using data that does not meet requirements for a specific purpose can result in ML models that are inaccurate and prone to failure. Therefore, to help organization ensure that data for analytics and ML meet requirements, the ISO/IEC 5259 series identifies data quality characteristics, data quality measures, data quality management requirements and a representative process to manage data quality over the data life cycle along with the concepts data record and data item for applying to data quality management, in addition to a governance framework to direct and oversee the implementation and operation of all that.

5.1.2 Machine learning and data quality

ISO/IEC 22989 defines ML as the process of optimizing model parameters through computational techniques such that the model's behaviour reflects the data or experience. ISO/IEC 23053 further describes ML as a branch of AI that employs computational techniques to enable systems to learn from data or experience. ML can perform diverse tasks using data and ML algorithms. The data used in ML are categorized as training data, validation data, testing data and production data. In supervised ML, an ML model is created by training an ML algorithm with training data. Validation data and testing data are then used to ensure the trained ML model performs in accordance with the organization's requirements. The trained ML model is then used to calculate inferences from production data. The performance of a trained ML model is dependent on the quality characteristics of all these types of data. ISO/IEC 23053 describes several general types of ML algorithms, which can have different sensitivities to different data quality characteristics.

EXAMPLE 1

Representativeness is one of the most important data quality characteristics for ML. When the training data does not represent the population included in the production data, the trained ML model has a higher probability of making incorrect inferences from the production data. When used to make decisions about people, this can lead to biased actions for underrepresented groups of people.

EXAMPLE 2

Training an ML algorithm to produce a trained ML model is a mathematical process that iterates over a set of training data that represents attributes of an object or event. The quality of each sample in the training data will influence the trained ML model. If too many samples in the training data are not accurate, the model is likely to produce incorrect inferences on production data.

NOTE See ISO/IEC 5259-2 for details of how data quality characteristics impact the performance of ML models.

5.1.3 Data characteristics that pose quality challenges for analytics and machine learning

Datasets which exhibit considerable variety or variability can influence the data quality model and associated data quality measures. Large volumes of data and data which are rapidly generated or changing can require the use of automated tools to perform data quality measures and to assess whether the data meet requirements. Large volumes of data can also create challenges for just-in-time data quality measurement and assessment.

5.1.4 Data sharing, data re-use and data quality for analytics and machine learning

The same data can be used for different analytics or ML projects. For example, data can be shared by a data holder with multiple data users (internal or external to the data holder's organization). Likewise, a data user can be allowed to use the data for more than one task.

Different analytics and ML projects can come with different data quality requirements. Different data quality requirements can affect the choice of a data quality model, associated data quality measures and assessment criteria.

5.2 Data quality concept framework for analytics and machine learning

5.2.1 Overview

Figure 1 provides a representative framework and relation with ISO/IEC 5259 series and ISO 8000-120 for determining, assessing and improving the quality of a dataset for use in analytics and ML. The aim of the framework in Figure 1 aims to identify processes that can be used to determine and ensure that the dataset meets the needs and requirements of the organization.

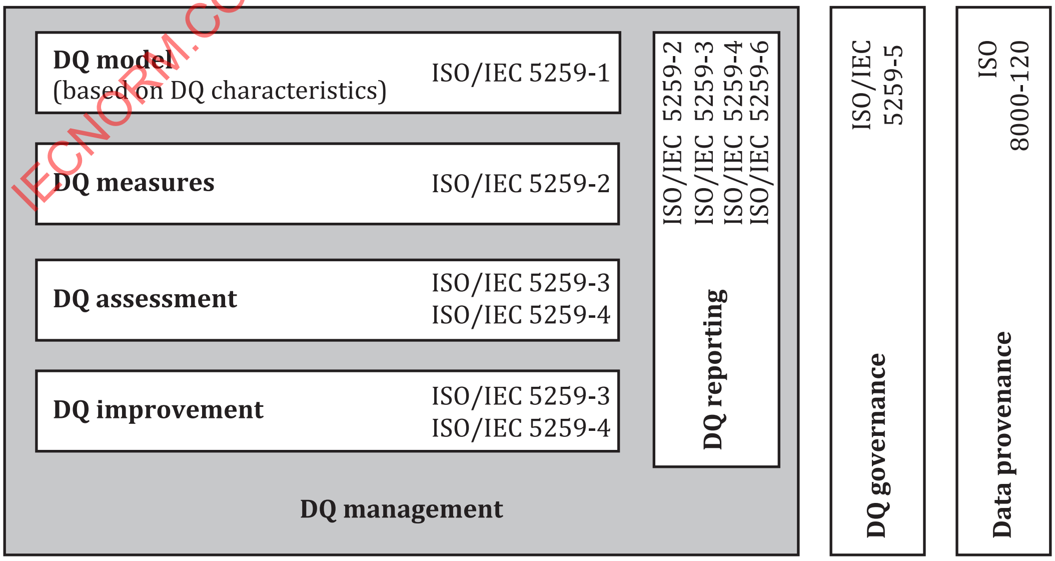


Figure 1 — Data quality concept framework for analytics and machine learning

Elements in the framework that are specific to data quality include the DQ model, DQ measures, DQ assessment, DQ improvement and DQ reporting. Other important processes include DQ governance, DQ management and data provenance.

DQ measures, DQ assessment and DQ improvement processes can be iterative when needed to meet organizations' needs and requirements for the dataset.

Additionally, for continuous learning (i.e. where the ML algorithm is continuously trained with new data) these processes can also be applied continuously over the system's life cycle.

5.2.2 Data quality management

5.2.2.1 Data quality model

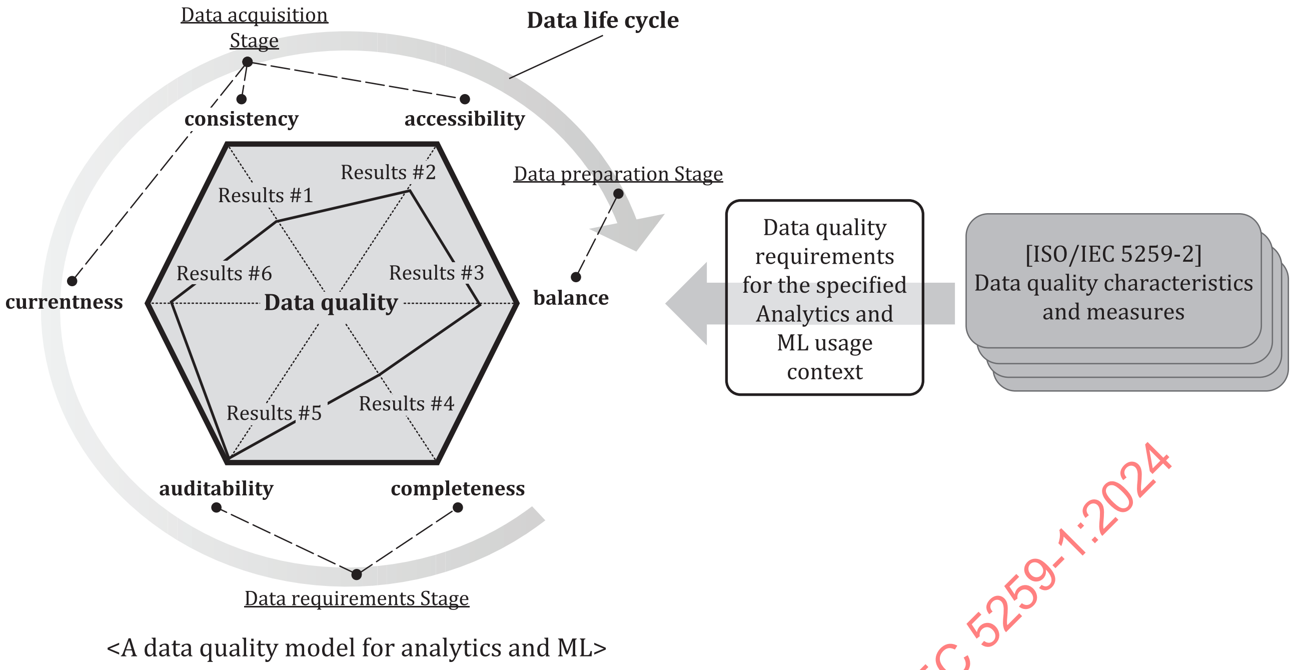
For the purposes of this document, a data quality model is a defined set of data quality characteristics which provides a framework for specifying data quality requirements and evaluating data quality. Data users can establish data quality models for analytics and ML according to their business objectives.

ISO/IEC 5259-2:—, Clause 6 outlines data quality characteristics, in accordance with inherent, system dependent points of view derived from ISO/IEC 25012 as well as additional data quality characteristics for analytics and ML.

ISO/IEC 25012 describes these two views of data quality as follows:

- Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data are used under specified conditions.
- System dependent data quality refers to the degree to which data quality is reached and preserved within a computer system when data are used under specified conditions.

Figure 2 shows the relationship of a data quality model for analytics and ML, a data life cycle (see 5.3), data quality requirements, and data quality characteristics for analytics and ML. Data users establish data quality models for analytics and ML according to their business objectives. A data quality model is composed of a defined set of measurable data quality characteristics. Measurement of data quality characteristics can be conducted during appropriate stages of the data life cycle.



Key

⟳ data life cycle for analytics and ML (see 5.3)

Figure 2 — Example of applying data quality characteristics from ISO/IEC 5259-2

In Figure 2, the data user has selected six data quality characteristics, based on data quality requirements for their data quality model out of all the data quality characteristics defined in ISO/IEC 5259-2. Appropriate data quality measures are then selected for each data quality characteristic to determine a value for each of them; an example of visualization of results is shown in the internal hexagon.

5.2.2.2 Data quality measures

Once the data quality model is defined according to the analytics and ML usage context, data users can select appropriate data quality measures to evaluate each of the data quality characteristics in the data quality model.

NOTE See ISO/IEC 5259-2 for more information on data quality measures as used in analytics and ML.

5.2.2.3 Data quality assessment

An organization can use the results of selected data quality measures to assess whether a dataset meets its needs and requirements.

If the organization determines that a dataset meets its needs and requirements, the dataset can then be used to train, validate and test an ML algorithm or operate a trained ML model or be used in analytics processes. The process to determine whether the dataset meets its requirements can occur in an iterative fashion as detailed in 5.3.

If the organization determines that the dataset does not meet its needs and requirements, the organization can choose to:

- attempt to improve the dataset;
- discontinue use of the dataset;
- acquire a new dataset.

5.2.2.4 Data quality improvement

Data transformations can be applied to a dataset to improve its quality to the extent it meets the needs and requirements of the organization. Data quality should be addressed as far upstream as possible (e.g. by data originators and data holders) to reduce the workload on data users and to provide more consistency in downstream versions of the dataset.

Data quality improvement should be considered in the context of the organization's business, data quality requirements and ML model performance. It is not always necessary to incur the time and expense to improve the data to a 100 % state for all data quality characteristics to meet requirements.

NOTE See ISO/IEC 5259-4 for more information on data quality improvement.

5.2.2.5 Data quality reporting

The organization can develop and publish data quality reports according to its internal policies. These reports can help determine the root cause(s) for poor performance of ML models and other analytical tasks, and can be helpful for transparency and explainability of ML. Data quality reports can include:

- intended use of the dataset;
- data quality thresholds relative to the organization's needs and requirements for the dataset;
- data quality characteristics selected for the data quality model;
- explanations for specific data quality characteristics not included in the data quality model;
- data quality measures used for each data quality characteristic;
- results of data quality measurements;
- data quality trends (e.g. whether data quality is observed to be improving or declining);
- actions taken to improve the quality of the dataset;
- assessment of whether the dataset meets the organization's needs and requirements;
- people involved in the data quality model, measurement and improvement processes.

Data visualization provides tools to explore data, as well as ways to effectively communicate the results of data quality processes. From data acquisition to data decommissioning stages, data visualization can make checking the status of data easier by using methods such as those described in ISO/IEC 20547-3:2020, 9.2.2.5. Especially in the data preparation stage, data visualization can help:

- with data cleaning by finding incorrect values, missing values and duplicate values;
- in creating and selecting attributes or features;
- in merging attributes or features as part of the data reduction process;
- to explain data trends, patterns, distributions and outliers;
- to show the relationship between data quality measures and established thresholds (e.g. using red, yellow and green indicators).

NOTE For more information on data quality visualization, see ISO/IEC TR 5259-6.

5.2.3 Data quality governance

Data quality processes should conform to the organization's data governance policies. A culture of accountability is critical to achieving data quality in an organization. From a data quality perspective, data governance can provide:

- a set of guiding principles established by an organization to actively manage and improve data quality;
- decision-making structures and accountabilities through which those with assigned data quality responsibilities are held to account;
- organizational roles and responsibilities to ensure data quality through repeatable processes.

NOTE For more information on data quality governance framework, see ISO/IEC 5259-5. Data governance is described in ISO/IEC 38505 series and ISO/IEC 38507. Data governance for big data is described in ISO/IEC 20547-3:2020, 8.4 and Annex A.

5.2.4 Data provenance

Data quality processes for analytics and ML can be complex, with multiple and iterative steps. Data provenance records can be used to gather and maintain data provenance information which can provide a basis for determining whether the data have been intentionally manipulated or altered. These records can help data users assess their trust in the data. Data provenance records can include:

- the source or origin of the data;
- all processes applied to the data;
- all changes applied to the data (e.g. statistical transformations, modification of data values);
- all organizations or individuals who have had custody of the data since their creation.

NOTE For more information on data provenance records, see ISO/IEC 8000-120.

5.3 Data life cycle for analytics and ML

5.3.1 Overview

Analytics and ML both make predictions based on data which the organization can use to make decisions. Therefore, analytics and ML are highly dependent on the characteristics of data and the characteristics of data quality in terms of a usage context. At the same time, analytics and ML also share the high-level data life cycle. As per 5.2.2, data quality requirements depend on the purpose of the analysis and ML, so data quality requirements should be managed according to the purpose and the life cycle of the data used.

The data life cycle for analytics and ML provides an end-to-end description of how data are used and additional data are generated within analytics or ML system.

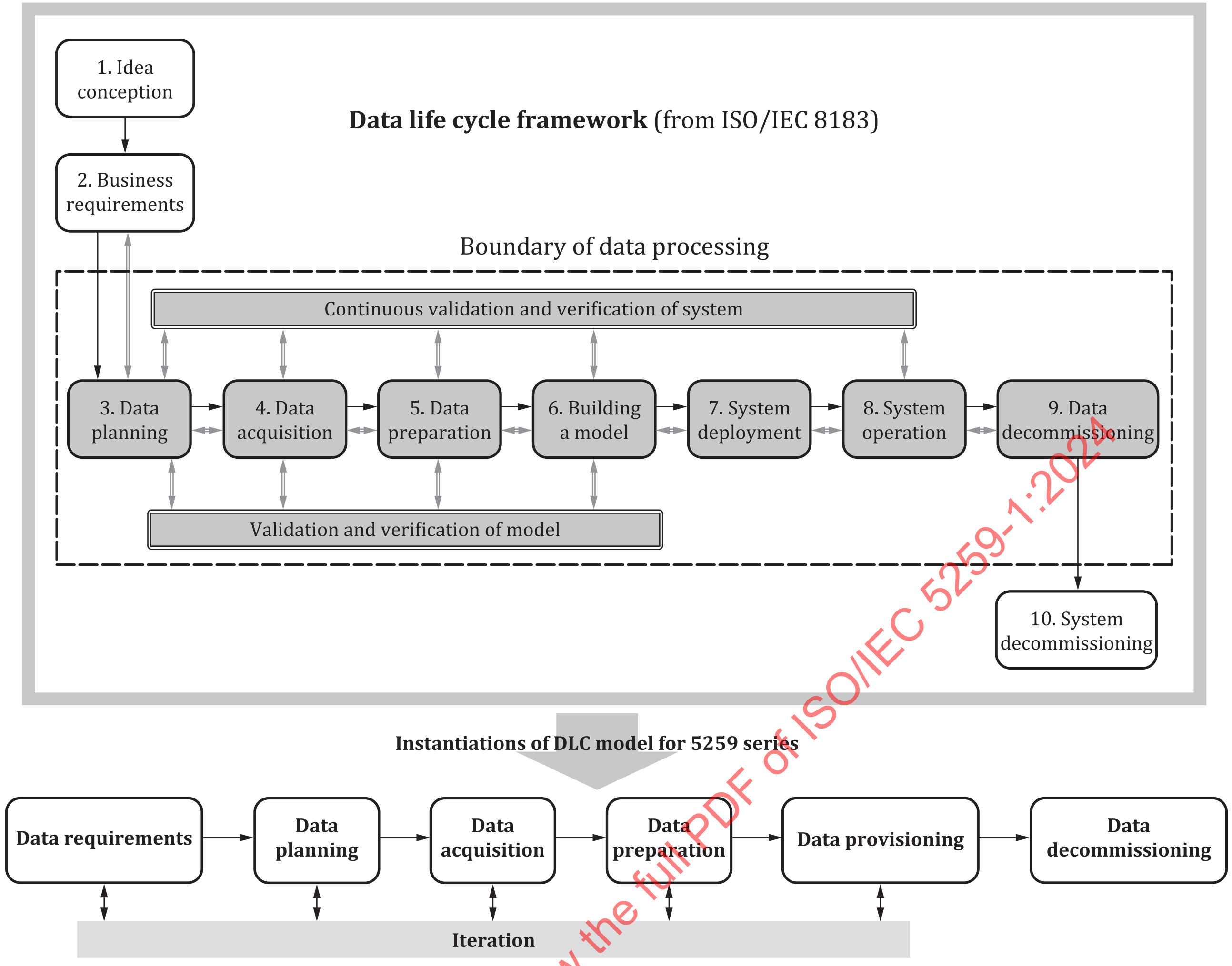
This document defines DLC for analytics and ML through:

- the six-stage DLC model (see 5.3.2);
- processes across the multiple stages of the DLC model (see 5.3.3).

5.3.2 Data life cycle model

5.3.2.1 Overall model

The DLC model for analytics and ML shown in Figure 3 is derived from ISO/IEC 8183 and identifies stages that provide context for the processes used in data quality management. The single-headed arrows represent the main progression through the stages and the double-headed arrows represent feedback to other stages of the DLC model.



Key

▭	data life cycle for analytics and ML (see 5.3)
▬	iteration
→	primary development pathway
↔	feedback pathway

Figure 3 — Data life cycle model for analytics and ML

5.3.2.2 Stage 1: data requirements

The data requirements stage involves:

- determining what data are required for an analytics or an ML project;
- checking the availability of the data for an analytics or an ML project;
- instantiating a data quality model with relevant data quality characteristics.

5.3.2.3 Stage 2: data planning

The data planning stage ensures that the data to be used meet the requirements of the data requirements stage while supporting the goals of the analytics and ML project that uses the data. This stage involves:

- designing the data architecture (i.e. defining the full nature and extent of the data needed and how the data will be used);

- estimating the effort required to acquire and prepare data to support the analytics or ML project, which can include any required restructuring of the data, time to transfer or collect the data and building a data quality model for the analytics or ML project.

5.3.2.4 Stage 3: data acquisition

The data acquisition stage involves collecting data that are used in the analytics or ML project. Live and historic data are collected from the sources identified in the data planning stage. Depending on the analytics or ML project requirements, this data can arrive as a stream or in batches. This stage involves:

- protecting the privacy of data subjects and securing the data;
- testing and if necessary, improving data collection methods;
- performing data quality measurements and if necessary, improving data quality. This potentially reduces the workload for data users and reduces the risk of introducing downstream inconsistencies in the data from the application of different transforms.

NOTE For more information on the data acquisition stage, see ISO/IEC 8183.

5.3.2.5 Stage 4: data preparation

In the data preparation stage, the collected data are processed into a form that can be used by the analytics and ML project. This stage plays an important role in meeting the data quality requirements and can be repeated according to the results of the analytics process or the performance of the trained ML model. This stage involves the following optional processes, depending on the identified data quality requirements:

- Transforming data: converting data from one representation or space to another with minimal loss of content.
- Validating data: ensuring that the data are correct based on the validation of data quality characteristics such as correctness, meaningfulness, security and privacy.
- Cleaning data: detecting inaccurate data or missing data and correcting the data by replacing, modifying, imputing or deleting.
- Aggregating data: combining two or more datasets into one dataset in summary form.
- Sampling data: selecting data from a dataset. Sampling can be done with or without replacement.
- Feature creation: creating new attributes that can capture the main information in data more efficiently than the original attributes.
- Feature selection: reducing the dimensionality of data by using a subset of the features available.
- Enrichment: linking diverse data sources and adding additional context to the data.
- Data labelling and annotation: training, validation and testing data for supervised ML requires values for one or more target variables. Data labelling is the process of assigning values for the target variables if they are not included in the acquired data.

NOTE 1 Different ML tasks can require additional, unique data preparation processes.

NOTE 2 For more information on data preparation, see ISO/IEC 8183 and ISO/IEC 5259-4.

5.3.2.6 Stage 5: data provisioning

In the data provisioning stage, the prepared data are applied to the analytics and ML project. In this stage, the performance of analytics or the trained ML model is assessed to determine if they meet requirements.

If the analysis results or performance of the ML model do not meet expectations, the following steps can be carried out:

- For analytics and ML, determine the extent to which training data or the algorithm can be the root cause.
- Communicate between the data originator and the data holder about data quality issues discovered at the data provisioning stage (e.g. data quality issues that adversely affect the performance of the ML model can be communicated to the data originator and data holder). Data originators and data holders can use this information to improve the quality of the data upstream to benefit future data users.
- Improve the data quality by repeating stages 2 to 4.

NOTE Sometimes the only way to achieve acceptable performance of an ML model is to use different data.

- Redo the analysis or rebuild the ML model.

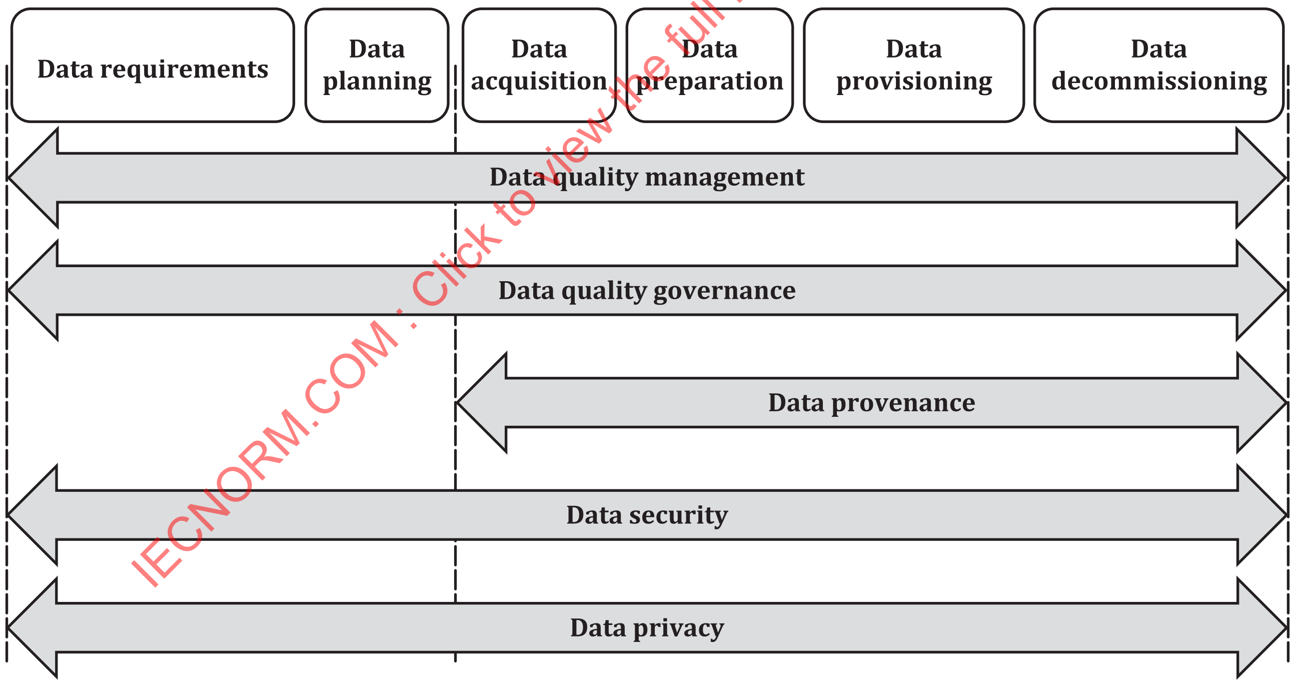
5.3.2.7 Stage 6: data decommissioning

During the data decommissioning stage, the data can be stored or archived with metadata for future use. In some cases, the data can be required to be destroyed or returned to a data holder. If the data are archived, the data requirements should be saved along with the data usage context for system-dependent data.

5.3.3 Processes across the multiple stages

5.3.3.1 General

Figure 4 shows processes that should apply across the multiple stages of the DLC model for analytics and ML. Data quality management, data quality governance and data provenance are described in 5.2.



Key

▭ stage where data are processed

⇔ data quality related tasks done in the lifecycle stage(s)

Figure 4 — Processes across the multiple stages

5.3.3.2 Data security

The dataset should be kept secure throughout all stages of the DLC model to ensure that it is available to authorized personnel and processes and that the data are not inappropriately altered. Inappropriate changes to the dataset can themselves cause incorrect results from ML models and other analytical tasks.

NOTE For more information on data security, see the ISO/IEC 27001 and related International Standards.

5.3.3.3 Data privacy

Datasets used for ML and analytics can contain PII, which should be protected in accordance with applicable requirements throughout all stages of the DLC model. De-identification techniques can be used to remove PII but production data used to make predictions about individuals can still contain or be linkable to PII.

NOTE For more information about data privacy, see ISO/IEC 27701.