International Standard

ISO/IEC 5259-4

First edition
2024-07

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

Part 4:
Data quality process framework

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

Partie 4: Cadre pour le processus de qualité des données

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/IEC 5259-4:2024
https://standards.iteh.ai/catalog/standards/iso/43cfc2a4-9441-4d50-9f28-3fe804cf02c9/iso-iec-5259-4-2024

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

Artificial intelligence (AI)-related products, systems or solutions have developed quickly in recent years. One of the common characteristics of an AI system, especially for systems using supervised machine learning (ML), is whether the AI system can be trained on a dataset before deployment or trained dynamically as the system is used.

Data have been recognized as one of the most important aspects of ML-based AI systems. For all supervised, semi-supervised, unsupervised and reinforcement learning approaches, data quality can be a primary concern in creating and using data for training and evaluating ML systems. It has been shown that with more accurate and richer data, the results of analytics and ML can be more useful and reliable. In addition, for the development of supervised learning-based AI systems, a large number of task-specific labelled training data is needed. This makes accurately labelled data one of the most important resources in the AI industry. Nowadays, there is a verified market of industrial services and tools for training data labelling. This market is now reaching a level of maturity that justifies the development of International Standards for the benefit of providers and users of these services and tools to ensure high-quality labelled data.

This document describes the implementation of a standardized common procedure of data processing with regard to data quality for analytics and ML. Clause 5 describes principles about data quality process and Clause 6 describes a data quality process framework. Clause 7 describes the data quality process for ML approaches, Clause 8 describes data labelling methods and process, Clause 9 provides roles of participants in data quality processes, Clauses 10 and 11 then describe the additional considerations that apply to semi-supervised learning and reinforcement learning. Clause 12 describes how the data quality process framework applies to analytics.

This document provides the process framework on a detailed level which can be used to fulfil the requirements specified in ISO/IEC 5259-3. It also links processes that are mapped on the data life cycle model in ISO/IEC 5259-1.

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

Part 4:
Data quality process framework

1 Scope

This document establishes general common organizational approaches, regardless of the type, size or nature of the applying organization, to ensure data quality for training and evaluation in analytics and machine learning (ML). It includes guidance 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.

This document is applicable to training and evaluation data that come from different sources, including data acquisition and data composition, data preparation, data labelling, evaluation and data use. This document does not define specific services, platforms or tools.

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, *Artificial intelligence — Data quality for analytics and machine leaning (ML) — Part 1: Overview, terminology and examples*

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *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 5259-1, 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
outsourcing
subcontracting of an activity by an organization to an external organization

[SOURCE: ISO TS 22002-4:2013, 3.14, modified — Definition revised.]

3.2
stand-off annotation
annotation layered over primary data and serialized in a document separate from that containing the primary data

[SOURCE: ISO 24612:2012, 2.7, modified — Note to entry removed.]

3.3
cloud service
one or more capabilities offered via cloud computing invoked using a defined interface

[SOURCE: ISO/IEC 22123-1:2023, 3.1.2]

3.4
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, personally identifiable information (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, the right to display a name, the right to identity and the right to prohibit data use in a way that offends honourable mention.

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

3.5
bounding box
rectangular region enclosing annotated object

[SOURCE: ISO/IEC 30137-4:2021, 3.3, modified — Note to entry removed.]

3.6
segmentation
process of separating the objects of interest from their surroundings

Note 1 to entry: Segmentation can be applicable to 2D, 3D, raster or vector data.

[SOURCE: ISO/IEC 3532-1:2023, 3.1.13]

3.7
key-point
point of interest on an object

4   Symbols and abbreviated terms

AI   artificial intelligence

AQL   acceptance quality limit

DLC   data life cycle

DQPF   data quality process framework

IoT   internet of things

JPEC   joint photographic experts group

JSON   JavaScript object notation

ML   machine learning

MPEG   moving picture experts group

PII   personally identifiable information

XML   extensible markup language

5   Data quality process principles

ISO/IEC 5259-1 defines data quality as a characteristic of data that the data meet the organization's requirements for a specified context.

Regardless of the data and evaluation methodology, the data quality process for analytics and ML should be based on general principles that are used throughout the DLC model. Organizations should define and document general data quality principles, taking into account the following aspects:

— the data and datasets are appropriate to the specified ML or analytics task;

— use of a data quality model based on data quality characteristics;

— validation of data quality according to data quality requirements using data quality measurements and specified targets;

— verifying at each stage that the process is on track to meet these targets and other requirements;

— correctness and robustness under testing, including techniques such as adversarial testing that are designed to identify errors;

— alignment with organizational requirements for security, privacy, fairness and ethics;

— protection of the health and well-being of annotators and other people involved in the data quality process's execution;

— documentation of progress and adherence to specified principles and requirements.

6   Data quality process framework

6.1   General

Based on the principles in Clause 5, the aim of the DQPF is to enable organizations to manage data quality to the extent the data meet requirements. Outcomes of the DQPF can include:

— data quality strategy;

— data quality plan;

— data quality requirements including a data quality model, data quality measures and targets;

— data quality work products (e.g. results of data quality measurements, failure reports, data improvement and augmentation methods applied);

— guidance on improvement to the data quality processes;

— approval for the use of the data in the analytics or ML project.

DLC perspective (as described in ISO/IEC 5259-1)

DLC

Data requirements → Data planning → Data acquisition → Data preparation → Data provisioning → Data decommissioning

Manage DLC with data quality products

Relevant organizations with DLC management

Data quality requirements as described in this document and ISO/IEC 5259-3

Data quality work products as described in ISO/IEC 5259-3

DQPF perspective (as described in ISO/IEC 5259-4)

Data quality planning
- Analyze data quality requirements
- Build a data quality model as described in ISO/IEC 5259-1 and 5259-2
- Determine the appropriate data quality measures as described in ISO/IEC 52529-2 and set target values

Data quality evaluation
- Apply data quality measures
- Compare data quality measures results against established targets
- Assess whether data requirements are met

Iteration

Data quality process validation
- Assessment of whether the data meets requirements
- If the data do not meet requirements, changes to the data quality processes are considered and implemented
- Complete data provenance records (if used)

Data quality improvement
- Apply data quality improvement methods as described in this document
- Repeat the data quality evaluation processes

Example outcomes of the DQPF
- Data quality strategy;
- Data quality plan;
- Data quality requirements including a data quality model, data quality measures and targets;
- Data quality work products (e.g. results of data quality measurements, failure reports, data improvement and augmentation methods applied);
- Guidance on improvement to the data quality processes;
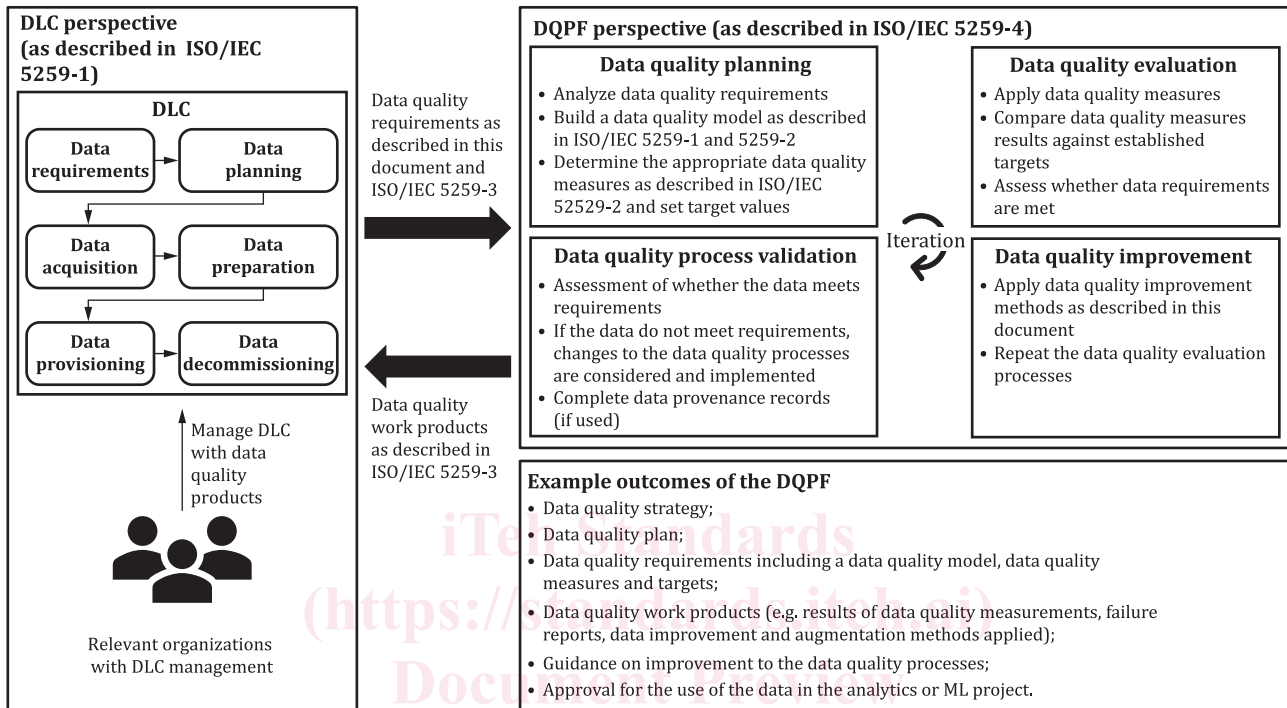- Approval for the use of the data in the analytics or ML project.

Figure 1 — The overall DQPF and examples of outcomes

Figure 1 provides a basic structure of the DQPF. The components of the DQPF include:

— data quality planning: establish the data quality management plans by analysing data quality requirements and data life cycle and determining the data quality management methods;

— data quality evaluation: measure and monitor data quality in the DLC model and provide results for the data quality plans;

— data quality improvement: implement data quality improvement processes (e.g. data cleaning, transforms, augmentation, scaling);

— data quality process validation: evaluate the data quality measures and processes to ensure the data meet requirements and provide feedback to the data quality improvement process if needed.

Figure 2 describes the relationship between the DLC model in ISO/IEC 5259-1 and the DQPF. The DQPF can be used throughout the overall DLC model to manage the data quality.
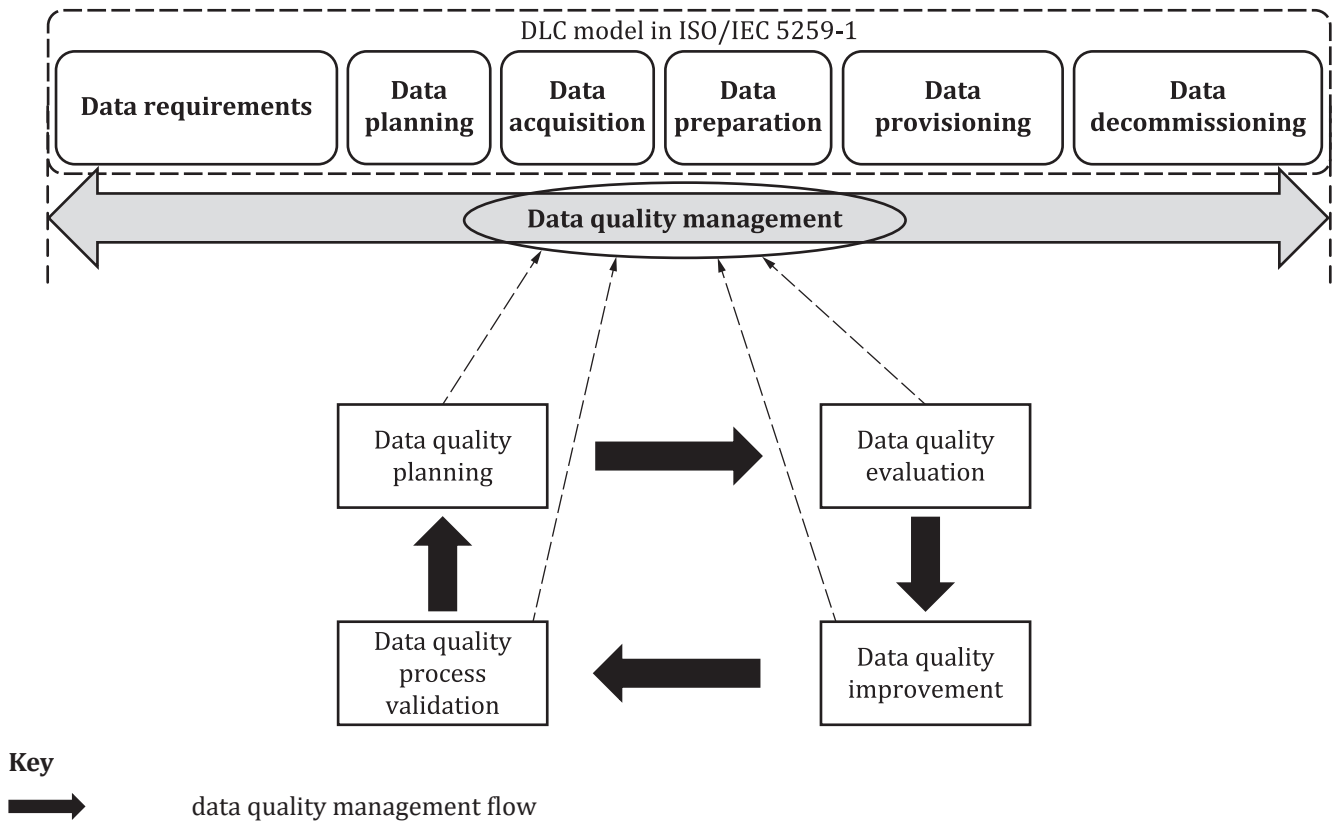
Key

➡ data quality management flow

iTeh Standards
(https://standards.iteh.ai)
Figure 2 — The relationship between the DLC model and the DQPF
Document Preview

6.2   Data quality planning

The data quality planning process activities and outcomes include:

ISO/IEC 5259-4:2024

— activities:
https://standards.iteh.ai/catalog/standards/iso/43cfc2a4-9441-4d50-9f28-3fe804cf02c9/iso-iec-5259-4-2024

— analyse the data quality requirements from the stakeholders in the data life cycle model;

— build the data quality model as described in ISO/IEC 5259-1 and ISO/IEC 5259-2;[1)]

— determine the appropriate data quality measures as described in ISO/IEC 5259-2 and target values for each measure;

— outcomes:

— data quality model;

— data quality measures;

— data quality measurements targets;

— data quality work products, such as results of data quality measurements, failure reports, data improvement and augmentation methods applied.

---

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