FINAL DRAFT
International Standard

ISO/IEC FDIS 5259-1

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

Part 1: Overview, terminology, and examples

ISO/IEC JTC **1**/SC **42**

Secretariat: **ANSI**

Voting begins on:
2024-04-01

Voting terminates on:
2024-05-27

Reference number
ISO/IEC FDIS 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[2)] 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[3)] 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[4)] 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[5)] 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 DIS 5259-2:2023.

2) Under preparation. Stage at the time of publication: ISO/IEC FDIS 5259-3:2024.

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

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

5) Under preparation. Stage at the time of publication: ISO/IEC WD 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

[SOURCE: ISO/IEC FDIS 38500:2023, 3.4, modified — The term “data governance” added as a preferred term, definition revised.]

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.]