



International
Standard

ISO/IEC 5259-2

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

**Part 2:
Data quality measures**

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

Partie 2: Mesure de la qualité des données

**First edition
2024-11**

Reference number
ISO/IEC 5259-2:2024(en)

Full document details: ISO IEC 05259-2-2024

 **Documents**
© ISO/IEC 2024 Document Sharing Hub



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2024

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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	v
Introduction	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Symbols and abbreviated terms	5
5 Data quality components and data quality models for analytics and machine learning	5
5.1 Data quality components in data life cycle	5
5.2 Data quality model	6
6 Data quality characteristics and quality measures	8
6.1 General	8
6.2 Inherent data quality characteristics	9
6.2.1 Accuracy	9
6.2.2 Completeness	10
6.2.3 Consistency	12
6.2.4 Credibility	13
6.2.5 Currentness	14
6.3 Inherent and system-dependent data quality characteristics	15
6.3.1 Accessibility	15
6.3.2 Compliance	15
6.3.3 Efficiency	16
6.3.4 Precision	16
6.3.5 Traceability	17
6.3.6 Understandability	17
6.4 System-dependent data quality characteristics	18
6.4.1 Availability	18
6.4.2 Portability	18
6.4.3 Recoverability	19
6.5 Additional data quality characteristics	19
6.5.1 Auditability	19
6.5.2 Balance	20
6.5.3 Diversity	22
6.5.4 Effectiveness	23
6.5.5 Identifiability	24
6.5.6 Relevance	25
6.5.7 Representativeness	25
6.5.8 Similarity	26
6.5.9 Timeliness	27
7 Implementing a data quality model and data quality measures for an analytics or ML task	28
8 Data quality reporting	28
8.1 Data quality reporting framework	28
8.2 Data quality measure information	29
8.3 Guidance to organizations	29
Annex A (informative) Design and document of a measurement function	30
Annex B (informative) UML model of data quality measure framework	32
Annex C (informative) Overview of data quality characteristics	33
Annex D (informative) Alternative groups of data quality characteristics	35

Annex E (informative) Comparison between data quality characteristics of ISO/IEC 25012 and ISO/IEC 5259-2.....36

Bibliography.....37

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-supported decision-making brings new challenges to data quality management in data analytics and artificial intelligence (AI) based on machine learning (ML). Issues in data quality, such as incomplete, false or outdated data, can adversely affect analytics and ML processes and outcomes. Data from various sources, including structured data (e.g. relational databases) and unstructured data (e.g. documents, images, audios), can be directly consumed into the data life cycle for analytics and ML model development. Data are transformed in each stage of the data life cycle of analytics and ML. A holistic standardized approach to control, produce and deliver sufficient high-quality data is necessary for data analytics and ML models to be safe, reliable and interoperable. To develop credible data quality management for analytics and ML, intrinsic data quality International Standards, including concepts and use cases, characteristics and measurements, management requirements, and process framework, can be considered.

This document is a part of the ISO/IEC 5259 series. This document builds upon the ISO 8000 series, ISO/IEC 25012 and ISO/IEC 25024. The purpose of this document is to describe a data quality model through the definition of data quality characteristics and data quality measures based on ISO/IEC 25012 and ISO/IEC 25024. Data quality models can be extended or modified according to this document.

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

Part 2: Data quality measures

1 Scope

This document specifies a data quality model, data quality measures and guidance on reporting data quality in the context of analytics and machine learning (ML).

This document is applicable to all types of organizations who want to achieve their data quality objectives.

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 learning (ML) — Part 1: Overview, terminology, and examples*

ISO/IEC 25024, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*

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 5259-1, ISO/IEC 22989 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

re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

Note 1 to entry: Data can be processed by humans or by automatic means.

[SOURCE: ISO/IEC 2382:2015, 2121272]

3.2

data frame

set of data records represented by a specific domain or purpose, with a shared structure of data items

Note 1 to entry: A data frame is two-dimensional, like a table with rows and columns. The term is specifically used in analytics and ML, e.g. in the R language, while other languages use “data set” to mean the same thing. In this document, “dataset” has a more generic meaning.

3.3

data type

categorization of an abstract set of possible values, characteristics, and set of operations for an attribute

Note 1 to entry: Examples of data types are character strings, texts, dates, numbers, images and sounds.

[SOURCE: ISO/IEC 25024:2015, 4.16]

3.4

data value

content of data item

Note 1 to entry: In ISO/IEC 25012:2008, 5.1.1, it is specified that from the inherent point of view, data quality refers to data itself such as data domain values and possible restrictions.

Note 2 to entry: Number or category assigned to an attribute of a target entity by making a measurement.

[SOURCE: ISO/IEC 25024:2015, 4.17]

3.5

empty data item

data item whose *data value* ([3.4](#)) has no value, i.e. Null or None

Note 1 to entry: This definition in general signifies non-existence of a data value (i.e. Null or None). A data item with string data type can be an empty data item by using either the empty string or Null. However, there is an exception for some application a string can be empty (e.g. “”) but not Null and hence not imply an empty data item.

3.6

entity

concrete or abstract thing in the domain under consideration

[SOURCE: ISO 8000-2:2022, 3.3.3]

3.7

raw data

data in its originally acquired, direct form from its source before subsequent processing

[SOURCE: ISO 5127:2017, 3.1.10.04]

3.8

target data

data ([3.1](#)) used in an analytics or ML task whose quality is measured

3.9

target population

population of interest in the analytics or ML project to which inferences are to be made

3.10

data quality subject

entity ([3.6](#)) affected by data quality

3.11

quality measure element

measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by a mathematical function

[SOURCE: ISO/IEC 25024:2015, 4.32]

3.12

quantity

property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference

[SOURCE: ISO/IEC Guide 99:2007, 1.1, modified — Notes to entry deleted.]

3.13

quantity value

number and reference together expressing magnitude of a *quantity* ([3.12](#))

[SOURCE: ISO/IEC Guide 99:2007, 1.9, modified — Examples deleted.]

3.14

measurement function

algorithm or calculation performed to combine one or more *quality measure elements* ([3.11](#))

[SOURCE: ISO/IEC 25021:2012, 4.7, modified — Definition revised.]

3.15

measurement result

result of measurement

set of *quantity values* ([3.13](#)) being attributed to a measurand together with any other available relevant information

[SOURCE: ISO/IEC Guide 99:2007, 2.9, modified — Notes to entry deleted.]

3.16

measure

<noun> variable to which a value is assigned as the result of measurement

Note 1 to entry: The plural form “measures” is used to refer collectively to base measures, derived measures and indicators.

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.15]

3.17

measure

<verb> make a measurement

[SOURCE: ISO/IEC 25000:2014, 4.19]

3.18

bounding box

rectangular region enclosing annotated object

Note 1 to entry: The major and minor axes of the rectangle are parallel to the edges of the images. For rotated boxes, the polygon annotation is to be used.

[SOURCE: ISO/IEC 30137-4:2021, 3.3]

3.19

cluster

automatically induced category of elements that are part of the dataset and that share common attributes

Note 1 to entry: Clusters do not necessarily have a name.

[SOURCE: ISO/IEC 23053:2022, 3.3.2]

3.20

clustering algorithm

algorithm which groups *clusters* (3.19) from input data

Note 1 to entry: Examples of clustering algorithms include centroid-based clustering, density-based clustering, distribution-based clustering, hierarchical clustering and graph-based clustering.

3.21

overfitting

<machine learning> creating a model which fits the training data too precisely and fails to generalize on new data

Note 1 to entry: Overfitting can occur because the trained model has learned from non-essential features in the training data (i.e. features that do not generalize to useful outputs), excessive noise in the training data (e.g. excessive number of outliers), a significant mismatch between training data and production data distributions or because the model is too complex for the training data.

Note 2 to entry: Overfitting can be identified when there is a significant difference between errors measured on training data and on separate test and validation data. The performance of overfitted models is especially impacted when there is a significant mismatch between training data and production data.

[SOURCE: ISO/IEC 23053:2022, 3.1.4]

3.22

fidelity

degree to which a model or simulation reproduces the state and behaviour of a real-world object or the perception of a real-world object, feature, condition, or chosen standard in a measurable or perceivable manner

[SOURCE: ISO 16781:2021, 3.1.4]

3.23

maintainability

ability of a functional unit, under given conditions of use, to be retained in, or restored to, a state in which it can perform a required function when maintenance is performed under given conditions and using stated procedures and resources

Note 1 to entry: The term used in IEC 191-02-07 is “maintainability performance” and the definition is the same.

Note 2 to entry: maintainability: term and definition standardized by ISO/IEC [ISO/IEC 2382-14:1997].

Note 3 to entry: 14.01.06 (2382)

[SOURCE: ISO/IEC 2382:2015, 2123027]

3.24

reliability

consistency with which an assessment measures

EXAMPLE An assessment will have low reliability if two assessment forms are of unequal difficulty or coverage or if there are errors in the scoring procedures or in the reporting of scores.

[SOURCE: ISO/IEC 23988:2007, 3.21]

3.25

validity

extent to which an assessment achieves its aim by measuring what it is supposed to measure and producing results which can be used for their intended purpose

Note 1 to entry: An assessment has low validity if the results are unduly influenced by skills which are irrelevant to the stated aims of the assessment.

[SOURCE: ISO/IEC 23988:2007, 3.25]

4 Symbols and abbreviated terms

AI	artificial intelligence
CSV	comma separated values
HDF	hierarchical data format
JSON	JavaScript object notation
ML	machine learning
IP	internet protocol
PII	personally identifiable information
QM	quality measure
UML	unified modelling language

5 Data quality components and data quality models for analytics and machine learning

5.1 Data quality components in data life cycle

[Figure 1](#) shows data quality components aligned with the data life cycle model shown in ISO/IEC 5259-1:2024, Figure 3, which can support data quality management processes. ISO/IEC 5259-1 defines a data quality model as a defined set of data quality characteristics. The data quality characteristic provides a framework for data quality requirements, implementation and evaluation methods. Data quality measures are variables assigned to which values are the results of measurements of data quality characteristics. Data quality measures are used to assess whether the data meet data quality requirements. Data quality measures can also be used to monitor and report data quality.

Target data are the data subject to data quality measurements. Target data can be raw data or data that has undergone one or more processes or transformations. Target data for measuring quality can be training, testing, validation, production and output data in the context of the use of analysis and ML (as described in ISO/IEC 23053).^[1] Target data can be formed as either data items or datasets. A data item consists of an item name, data value and data type representing a domain of values (e.g. character strings, texts, dates, numbers, images, sounds). A dataset can be classified into three forms:

- a collection of data items;
- a collection of data records;
- a collection of data frames.

The target data can be unlabelled or labelled depending on the association with data labels in the use of analytics or ML task.

NOTE This document makes no distinction between data structures, such as structured data, semi-structured data and unstructured data, or data roles, such as master data, transaction data and reference data.

Data quality reports are documents that express data quality requirements, the data quality model of data quality characteristics, data quality measures, the results of data quality measurements and an assessment of whether the data meet data quality requirements.

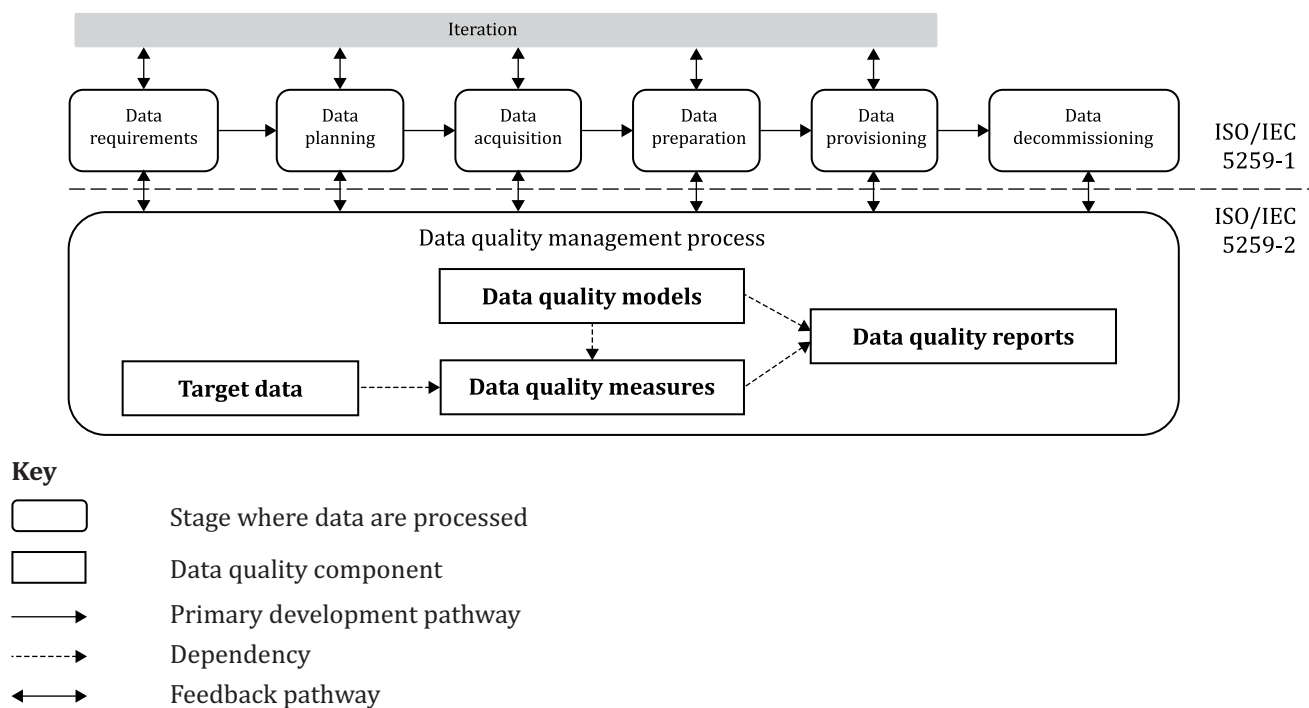


Figure 1 — Data quality components in data life cycle for analytics and ML

5.2 Data quality model

The data quality model provides a framework for specifying data quality requirements and evaluating data quality. In practice, a data quality model brings together data quality subjects, data quality characteristics and data quality requirements, for the context of the use of the data. The organization can specify data quality models by selecting data quality characteristics and measures to achieve target quality requirements for target data. [Figure 2](#) provides a UML diagram of the relationships between the components of the data quality model.

A data usage scope describes how and where the data can be used in an analytics or ML task and how it fits into an AI system.

EXAMPLE The data can be used to train a deep neural network ML model to predict product sales based on the features of a marketing strategy. The model can be trained and deployed using cloud services.

A data quality subject represents an entity affected by data quality. A data quality characteristic is a category of data quality attributes that bear on data quality (e.g. accuracy, completeness, precision). A data quality requirement describes properties or attributes of the data along with acceptance criteria relative to the data usage scope. Acceptance criteria can be quantitative or qualitative.

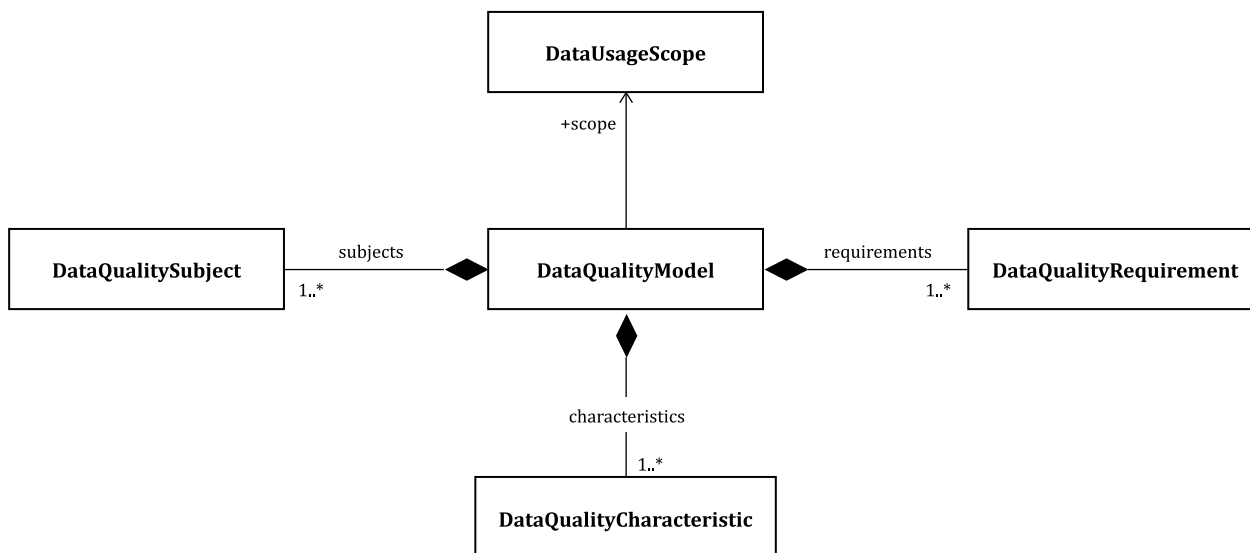


Figure 2 — Data quality model

When one quality characteristic affects another, trade-offs can be made by evaluating each requirement regarding importance and impact. In addition, it is crucial to balance the cost of data quality management with the priority of data quality requirements in determining how data quality characteristics and measures are incorporated into the data quality model. The organization can select the data quality characteristics and measures that correspond to their needs and requirements. Data quality should be assessed by comparing the results of selected data quality measures against established targets as established by data requirements. Any failures to achieve data quality requirements should be mitigated. ISO/IEC 5259-3^[2] describes the requirements and recommendations of a data quality management system to be applied by the organization.

ISO 8000-8^[3] and ISO/IEC 25012^[4] describe data quality models. ISO 8000-8 defines three data quality characteristics as being syntactic (format), semantic (meaning), and pragmatic (usefulness) to support industrial data generally as a product of business and manufacturing processes. ISO/IEC 25012 defines a general data quality model for data retained in a structured format within a computer system as a part of a software product. ISO/IEC 25012 takes into account all data types (e.g. characters, strings, texts, dates, numbers, images, sounds). ISO/IEC 25012 provides fifteen data quality characteristics: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability and recoverability.

The ISO 8000 series^[5] addresses various aspects of data quality such as data governance, data quality management (including processing) and maturity assessment. The ISO/IEC 25000 series^[6] addresses product (software, systems, data, services) quality requirements and evaluation. This document describes how the data quality characteristics of ISO/IEC 25012 can be applied to a data quality model for analytics and ML. Furthermore, this document defines additional characteristics that can contribute to higher-quality ML models and applications, as shown in Figure 3. Organizations should use the data quality characteristics and data quality measures described in this document whenever possible. However, the data quality characteristics in this document cannot comprehensively cover aspects that support all organizations' needs regarding data quality. Organizations may design their own data quality model by extending the data quality characteristics and data quality measures to fit their data requirements.

NOTE 1 See Annex A for information on designing and documenting measurement functions.

NOTE 2 See Annex E for a comparison between the data quality characteristics in ISO/IEC 25012 and those in this document.

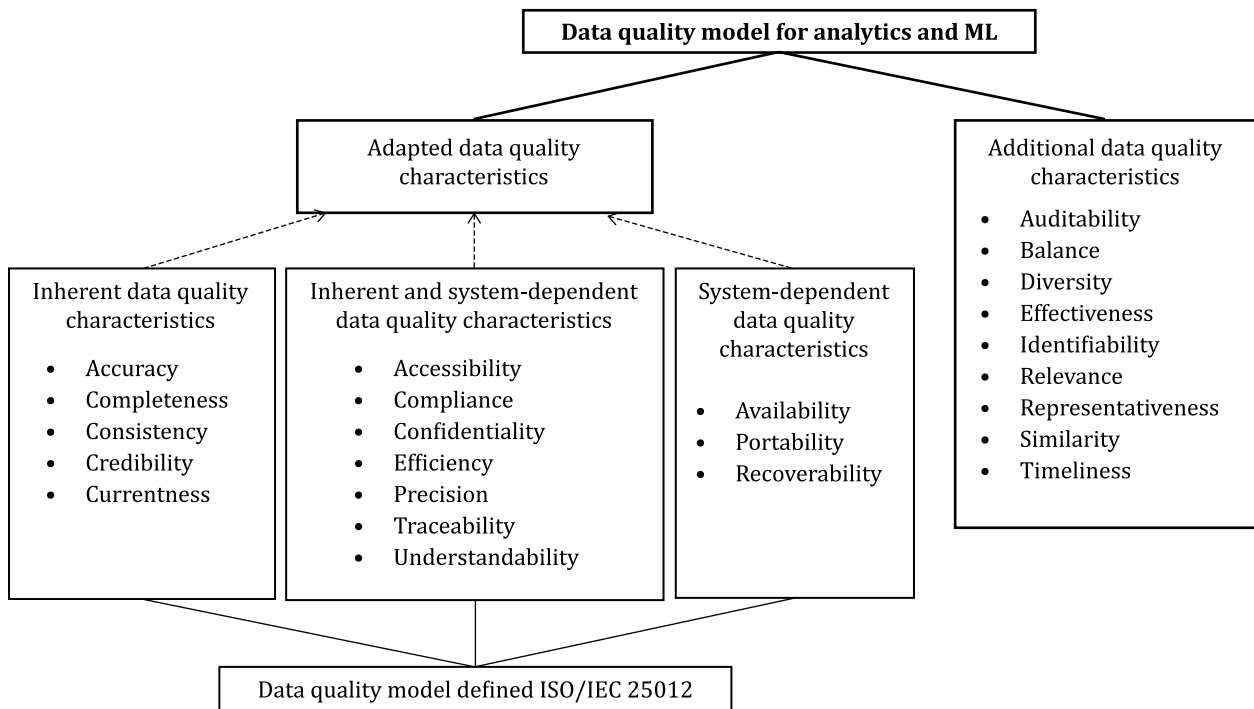


Figure 3 — Data quality characteristics for analytics and ML

6 Data quality characteristics and quality measures

6.1 General

Data quality characteristics and measures are used to specify and verify data quality requirements for identified attributes for target data. Each data quality characteristic is associated with one or more data quality measures for quantification. A data quality measure is a variable to which a value is assigned as the result of a measurement function. The data quality measures in this document are selected based on the context of use of analytics and ML.

NOTE 1 [Annex B](#) shows a framework for providing common vocabularies and relationships between the components of data quality measures.

NOTE 2 [Annex C](#) and [Annex D](#) show how quality measures are grouped from different perspectives.

In the context of analytics and ML, the overall quality of a training dataset, a validation dataset or a test dataset can be just as important as the quality of the individual data values in the dataset. Even though every data value in a dataset is accurate, a dataset that does not correctly reflect the underlying distribution of data can cause an incorrect analysis result or the creation of an ML model that does not meet requirements. The organization should document the target data for each data quality measure.

NOTE 3 Characteristics for statistical measures (e.g. accessibility by authorized users, accuracy, consistency, currentness, understandability, relevance, timeliness)^[7] as defined by institutions such as the United Nations Statistics Division (UNSD) and European Statistics (EUROSTAT) can also be used to assess whether the quality of a dataset meets requirements.

The data quality measures and measurement functions in this document should be used when appropriate. Refer to [Annex A](#) in cases where the user of this document needs to create a new, bespoke data quality measure and data quality measurement function. Any quality measure, when using modified or newly defined, shall select data quality characteristics defined in this document and shall provide the rationale for changes in accordance with ISO/IEC 25024:2015, Clause 2.