



DPC: 23/30432295 DC

BSI Group Headquarters

389 Chiswick High Road London W4 4AL

Tel: +44 (0)20 8996 9000

Fax: +44 (0)20 8996 7400

www.bsigroup.com

Date: 03 August 2023

Origin: International

Latest date for receipt of comments: 26 September 2023

Project No. 2021/00048

Responsible committee: ART/1 Artificial Intelligence

Interested committees:

Title: Draft BS ISO/IEC 5259-2 Artificial intelligence - Data quality for analytics and machine learning (ML). Part 2: Data quality measures.

Please notify admin.start@bsigroup.com if you are aware of any keywords that might assist in clarifying or identifying the standard or of the content of this standard:

- i) has any issues related to 3rd party IPR, patent or copyright
- ii) affects other national standard(s)
- iii) requires additional national guidance or information

WARNING: THIS IS A DRAFT AND MUST NOT BE REGARDED OR USED AS A BRITISH STANDARD. THIS DRAFT IS NOT CURRENT BEYOND 26 September 2023

This draft is issued to allow comments from interested parties; all comments will be given consideration prior to publication. No acknowledgement will normally be sent. **See overleaf for information on the submission of comments.**

No copying is allowed, in any form, without prior written permission from BSI except as permitted under the Copyright, Designs and Patent Act 1988 or for circulation within a nominating organization for briefing purposes. Electronic circulation is limited to dissemination by e-mail within such an organization by committee members.

Further copies of this draft may be purchased from BSI Shop <http://shop.bsigroup.com> or from BSI Customer Services, Tel: +44(0) 20 8996 9001 or email cservices@bsigroup.com. British, International and foreign standards are also available from BSI Customer Services.

Information on the co-operating organizations represented on the committees referenced above may be obtained from <https://standardsdevelopment.bsigroup.com/>.

E-mail: admin.start@bsigroup.com

Introduction

This draft standard is based on national and international discussions. Your comments on this draft are invited and will assist in the preparation of the consequent standard.

For international standards, comments will be reviewed by the relevant UK national committee before sending the consensus UK vote and comments to the international committee, which will then decide appropriate action. If the international standard is approved, it is usual for the text to be published as a British Standard.

For national standards, comments will be reviewed by the relevant UK national committee and the resulting standards published as a British Standard.

UK Vote

Please indicate whether you consider the UK should submit a negative (with supporting technical reasons) or positive vote on this draft. Please indicate if you are aware of any reason why this draft standard should not be published as a British Standard.

Submission of Comments

- **Annotated drafts are not acceptable and will be rejected.**
- All comments should be submitted online at <https://standardsdevelopment.bsigroup.com/>. You will need to register in order to comment.

Template for comments and secretariat observations

Date: xx/xx/20xx	Document: ISO/DIS xxxx
------------------	------------------------

1	2	(3)	4	5	(6)	7
MB	Clause No./ Subclause No./Annex (e.g. 3.1)	Paragraph/ Figure/Table/Note	Type of comment	Comment (justification for change) by the MB	Proposed change by the MB	Secretariat observations on each comment submitted
	3.1	Definition 1	ed	Definition is ambiguous and needs clarifying.	Amend to read '...so that the mains connector to which no connection...'	
	6.4	Paragraph 2	te	The use of the UV photometer as an alternative cannot be supported as serious problems have been encountered in its use in the UK.	Delete reference to UV photometer.	

DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 5259-2

ISO/IEC JTC 1/SC 42

Secretariat: ANSI

Voting begins on:
2023-08-03

Voting terminates on:
2023-10-26

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

Part 2: Data quality measures

ICS: 35.020

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.



Reference number
ISO/IEC DIS 5259-2:2023(E)
Documents
AVA Document Sharing Hub
© ISO/IEC 2023



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2023

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

31 Contents

32	Foreword.....	v
33	Introduction.....	vi
34	1 Scope.....	1
35	2 Normative references	1
36	3 Terms and definitions	1
37	4 Symbols and abbreviated terms	4
38	5 Data quality elements and data quality models for analytics and ML	4
39	5.1 Data quality elements in data life cycle	4
40	5.2 Data quality model.....	5
41	6 Data quality characteristics and quality measures	7
42	6.1 General.....	7
43	6.2 Inherent data quality characteristics.....	8
44	6.2.1 Accuracy	8
45	6.2.2 Completeness	9
46	6.2.3 Consistency	11
47	6.2.4 Credibility.....	11
48	6.2.5 Currentness.....	12
49	6.3 Inherent and system-dependent data quality characteristics.....	13
50	6.3.1 Accessibility	13
51	6.3.2 Compliance.....	14
52	6.3.3 Efficiency.....	14
53	6.3.4 Precision	14
54	6.3.5 Understandability.....	15
55	6.4 System-dependent data quality characteristics	16
56	6.4.1 Portability.....	16
57	6.5 Additional data quality characteristics	16
58	6.5.1 Auditability	16
59	6.5.2 Identifiability.....	17
60	6.5.3 Effectiveness	17
61	6.5.4 Balance	18
62	6.5.5 Diversity.....	20
63	6.5.6 Relevance.....	22
64	6.5.7 Representativeness.....	22
65	6.5.8 Similarity	23
66	6.5.9 Timeliness.....	24
67	7 Implementing a data quality model and data quality measures for an analytics or ML	
68	task.....	25
69	8 Data quality reporting.....	25
70	8.1 Data quality reporting framework.....	25
71	8.2 Data quality measure information	26
72	8.3 Guidance to organizations	26
73	Annex A (informative) Design and documenting a measurement function.....	27
74	Annex B (informative) UML model of data quality measure framework.....	29
75	Annex C (informative) Overview of data quality characteristics.....	30
76	Annex D (informative) Alternative categories of data quality characteristics	32

7 **Annex E (informative) Quality measures for synthetic data 34**

8 **Annex F (informative) Comparison between data quality characteristics ISO/IEC 25012 and**

9 **ISO/IEC 5259-2..... 35**

10 **Bibliography..... 37**

11

82 Foreword

83 ISO (the International Organization for Standardization) and IEC (the International Electrotechnical
84 Commission) form the specialized system for worldwide standardization. National bodies that are
85 members of ISO or IEC participate in the development of International Standards through technical
86 committees established by the respective organization to deal with particular fields of technical activity.
87 ISO and IEC technical committees collaborate in fields of mutual interest. Other international
88 organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the
89 work. In the field of information technology, ISO and IEC have established a joint technical committee,
90 ISO/IEC JTC 1.

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

95 Attention is drawn to the possibility that some of the elements of this document may be the subject of
96 patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any
97 patent rights identified during the development of the document will be in the Introduction and/or on
98 the ISO list of patent declarations received (see www.iso.org/patents).

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

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

105 This document was prepared by Technical Committee ISO/IEC JTC1, *Information technology*,
106 Subcommittee SC 42, *Artificial Intelligence*.

107 A list of all parts in the ISO/IEC 5259 series can be found on the ISO.org website.

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

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, should be considered.

This document is a part of the ISO/IEC 5259 series. 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 in accordance with this document.

Other parts of the ISO/IEC 5259 series include:

- ISO/IEC 5259-1:—¹⁾, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples*;
- ISO/IEC 5259-3:—²⁾, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines*;
- ISO/IEC 5259-4:—³⁾, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*.
- ISO/IEC 5259-5:—⁴⁾, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance*.

¹⁾ Under preparation. Stage at the time of publication: ISO/IEC CD 5259-1:2022

²⁾ Under preparation. Stage at the time of publication: ISO/IEC CD 5259-3:2022

³⁾ Under preparation. Stage at the time of publication: ISO/IEC CD 5259-4:2022

⁴⁾ Under preparation. Stage at the time of publication: ISO/IEC WD 5259-5:2022

Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

1 Scope

This document provides a data quality model, data quality measures and guidance on reporting data quality in the context of analytics and machine learning (ML). This document builds upon the ISO 8000 series, ISO/IEC 25012 and ISO/IEC 25024.

The aim of this document is to enable organizations to achieve their data quality objectives and is applicable to all types of organizations.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 20546:2019, ISO/IEC 22989:2022, ISO/IEC 5259-1: — and the following apply.

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

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

— IEC Electropedia: available at <http://www.electropedia.org/>

3.1

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]

3.2

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.

[ISO/IEC 2382:2015, 2121272]

3.3

dataset

collection of data with a shared format

EXAMPLE 1: Micro-blogging posts from June 2020 associated with hashtags #rugby and #football.

EXAMPLE 2: Macro photographs of flowers in 256x256 pixels.

Note 1 to entry: Datasets can be used for validating or testing an AI model. In a machine learning context, datasets can also be used to train a machine learning algorithm.

[SOURCE: ISO/IEC 22989:2022, 3.2.5]

3.4

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]

3.5

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]

3.6

data value

content of *data item* (3.4)

Note 1 to entry: In ISO/IEC 25012, 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]

3.7

data record

set of related *data items* (3.4) treated as a unit

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

3.8

data frame

set of *data records* (3.7) represented by a specific domain or purpose with the same *data items* (3.4), used for analytics and ML

3.9

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]

3.10

empty data item

data item (3.4) whose *data value* (3.6) has no value, i.e. Null or None

Note 1 entry: This definition in general signifies non-existence of a data value (i.e. Null or None), including for string data type. However, there is an exception for string data items, where string can be empty (e.g. “”) but not Null.

3.11

information

<information processing> knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning

[SOURCE: ISO/IEC 2382:2015, 212127]

- 214 **3.12**
 215 **measure**, noun
 216 variable to which a value is assigned as the result of measurement
- 217 Note 1 to entry: The plural form “measures” is used to refer collectively to base measures, derived measures and
 218 indicators.
- 219 [SOURCE: ISO/IEC 15939:2017, 3.15]
- 220 **3.13**
 221 **measure**, verb
 222 make a measurement
- 223 [SOURCE: ISO/IEC 25000:2014]
- 224 **3.14**
 225 **measured value**
 226 quantity value representing a measurement result
- 227 [SOURCE: JCGM 200:2012, 2.10]
- 228 **3.15**
 229 **measurement function**
 230 algorithm or calculation performed to combine two or more quality measure elements
- 231 [SOURCE: ISO/IEC 25021:2012, 4.7]
- 232 **3.16**
 233 **overfitting**
 234 <machine learning> creating a model which fits the training data too precisely and fails to generalize on
 235 new data
- 236 Note 1 to entry: Overfitting can occur because the trained model has learned from non-essential features in the
 237 training data (i.e. features that do not generalize to useful outputs), excessive noise in the training data (e.g.
 238 excessive number of outliers), a significant mismatch between training data and production data distributions or
 239 because the model is too complex for the training data.
- 240 Note 2 to entry: Overfitting can be identified when there is a significant difference between errors measured on
 241 training data and on separate test and validation data. The performance of overfitted models is especially impacted
 242 when there is a significant mismatch between training data and production data.
- 243 [SOURCE: ISO/IEC 23053:2022, 3.1.4]
- 244 **3.17**
 245 **quantity**
 246 property of a phenomenon, body, or substance, where the property has a magnitude that can be
 247 expressed as a number and a reference
- 248 [SOURCE: JCGM 200:2012, 1.1]
- 249 **3.18**
 250 **raw data**
 251 data in its originally acquired, direct form from its source before subsequent processing
- 252 [SOURCE: ISO 5127:2017, 3.1.10.04]
- 253 **3.19**
 254 **target data**
 255 data used in an analytics or ML task whose quality is measured

4 Symbols and abbreviated terms

AI	artificial intelligence
CSV	comma separated values
ML	machine learning
IP	internet protocol
PII	personally identifiable information
QM	quality measure
UML	unified modelling language

5 Data quality elements and data quality models for analytics and ML

5.1 Data quality elements in data life cycle

Figure 1 shows data quality elements aligned with the data life cycle model shown in ISO/IEC 5259-1:—, Figure 3, which can support data quality management processes. ISO/IEC 5259-1 defines a data quality model as a defined set of characteristics that provides a framework for data quality requirements and for evaluating data quality. Data quality measures are defined in ISO/IEC 5259-1 as variables to which values are assigned that are the results of measurements of data quality characteristics. Data quality measures are used to assess whether the data meet data quality requirements based on defined targets or thresholds. 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 transforms. Target data for ML tasks can be training, testing, validation, production and output data (as described in ISO/IEC 23053:2022). Target data can be formed as either data items or datasets. Datasets can be classified into four forms:

- a collection of data items;
- a collection of data records;
- a collection of data frames;
- a series of datasets sharing common characteristics.

In this document, target data are categorized into training data, validation data, test data, production data, and output data in the context of analysis and ML. 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). Datasets in this document are classified into four forms: a collection of data items, a collection of data records, a collection of data frames and a series of datasets sharing the same specification, such as structures and semantics. The target data can be unlabelled or labelled depending on the association with data labels in analytics or ML task.

NOTE This document makes no separation among 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, data quality measures' the results of data quality measurements and an assessment of whether the data meets data quality requirements.

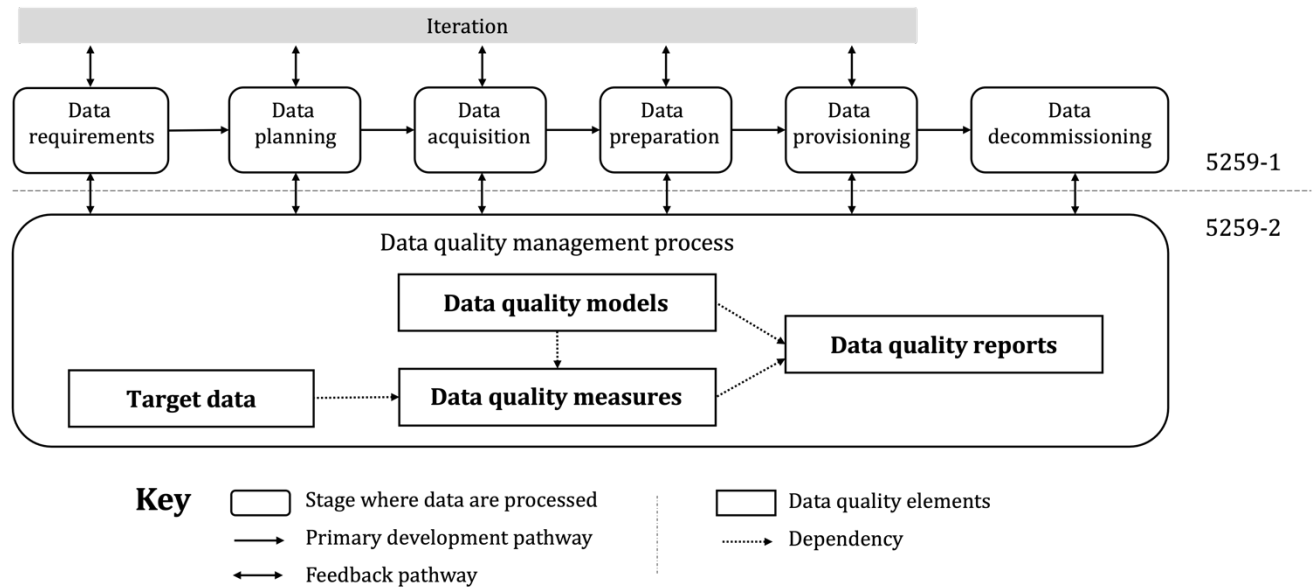


Figure 1 — Data quality elements in data life cycle for analytics and ML

5.2 Data quality model

ISO/IEC 5259-1 defines the concept of data quality model as a set of characteristics that 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 that are appropriate for the context of the use of the data. An organization can specify data quality models by selecting data quality characteristics and measures to archive target quality requirements for target data. Figure 2 provides a UML⁵ diagram of the relationships between these items.

Data usage context describes how and where the data will be used in an analytics or ML task and how it fits into system.

EXAMPLE

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

A data quality subject represents an entity affected by data quality. ISO/IEC 5259-1 defines data quality characteristic as 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 context. Acceptance criteria can be quantitative, qualitative or descriptive expressions.

⁵ UML is a standardized graphical language for visualizing, specifying, constructing and documenting the artifacts of software systems (see ISO/IEC 19505-1 [14]).

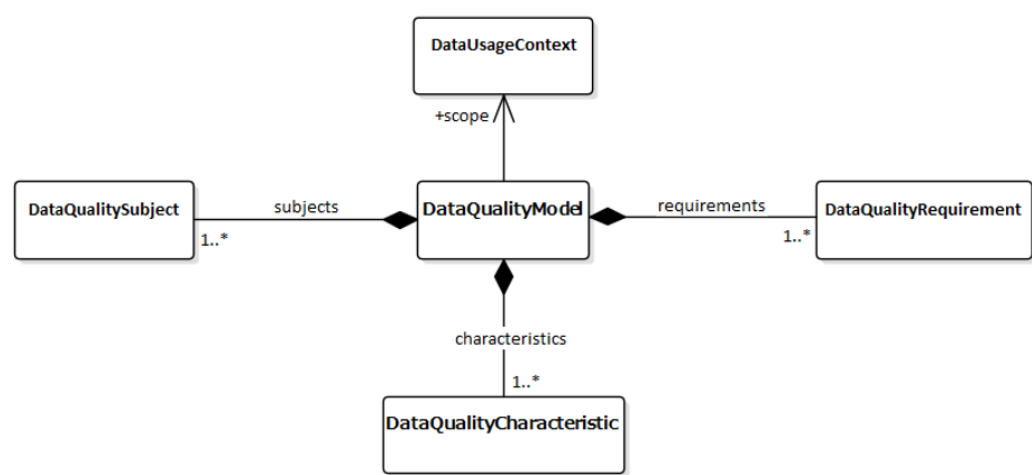


Figure 2 — Data quality model

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 quality of data should be assessed based on the organization’s needs and requirements, and any issues that result from data quality should be mitigated or otherwise addressed. The organization can select some or all of quality characteristics and measures that correspond to their needs and requirements. The quality of data should be assessed based on the selected quality characteristics and measures. Any issues that result from data quality should be mitigated or otherwise addressed.

ISO 8000-8 [12] and ISO/IEC 25012 [20] describe data quality models. ISO 8000-8 defines three data quality characteristics as being syntactic (format), semantic (meaning) and pragmatic (usefulness) to generally support industrial data 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 software product. ISO/IEC 25012 takes into account all data types of characters, strings, texts, dates, numbers, images, sounds, etc. ISO/IEC 25012 provides fifteen data quality characteristics: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability and recoverability.

ISO 8000 series [10] addresses various aspects of data quality such as data governance, data quality management (including processing) and maturity assessment. ISO/IEC 25000 series [18] addresses product (software, systems, data, services) quality requirements and evaluation. ISO/IEC 25000 series is outside the scope of the ISO 8000 series (see ISO 8000-1:2022 [11], 4). ISO/IEC 25000 series provides a practical view of data quality in the context of individual systems. This view complements the ISO 8000 series (see ISO 8000-1:2022, 8.4). This document adapts eleven data quality characteristics (five inherent, five inherent and system-dependent, and one system-dependent), which are shown with the underline in Figure 3, from fifteen characteristics defined in ISO/IEC 25012 for the data quality model and shows how they can be applied for analytics and ML. In addition, additional characteristics that can contribute to higher quality ML models and applications are defined as shown in Figure 3. However, the data quality characteristics in this document cannot comprehensively cover aspects to support AI quality. The organizations can define their own characteristics if necessary. See also Annex A for related measures. Annex F shows comparisons of data quality characteristics between ISO/IEC 25012 and this document.