भारतीय मानक

Indian Standard

IS/ISO/TS 8000-81 : 2021

आंकड़ों की गुणवत्ता

भाग 81 आंकड़ा गुणवत्ता आकलन — रूपरेखा

Data Quality

Part 81 Data Quality Assessment — Profiling

ICS 25.040.40

भारतीय मानक ब्यूरो
BUREAU OF INDIAN STANDARDS
मानक भवन, 9 बहादुर शाह ज़फर मार्ग, नई दिल्ली - 110002
MANAK BHAVAN, 9 BAHADUR SHAH ZAFAR MARG
NEW DELHI - 110002
www.bis.gov.in www.standardsbis.in

March 2024

Price Group 8

Industrial Automation Systems and Robotics Sectional Committee, PGD 18

NATIONAL FOREWORD

This Indian Standard (Part 81) which is identical to ISO/TS 8000-81 : 2021 'Data quality — Part 81: Data quality assessment — Profiling' issued by the International Organization for Standardization (ISO) was adopted by the Bureau of Indian Standards on recommendation of the Industrial Automation Systems and Robotics Sectional Committee and approval of the Production and General Engineering Division Council.

Other parts in this series are:

Part 1	Overview
Part 2	Vocabulary
Part 8	Information and data quality: Concepts and measuring
Part 60	Data quality management: Overview
Part 61	Data quality management: Process reference model
Part 62	Data quality management: Organizational process maturity assessment: Application of standards relating to process assessment
Part 63	Data quality management: Process measurement
Part 64	Data quality management: Organizational process maturity assessment: Application of the test process Improvement method
Part 65	Data quality management: Process measurement questionnaire
Part 66	Data quality management: Assessment indicators for data processing in manufacturing operations
Part 82	Data quality assessment: Creating data rules
Part 100	Master data: Exchange of characteristic data: Overview
Part 110	Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification
Part 115	Master data: Exchange of quality identifiers: Syntactic, semantic and resolution requirements
Part 116	Master data: Exchange of quality identifiers: Application of ISO 8000-115 to authoritative legal entity identifiers
Part 120	Master data: Exchange of characteristic data: Provenance
Part 130	Master data: Exchange of characteristic data: Accuracy
Part 140	Master data: Exchange of characteristic data: Completeness
Part 150	Data quality management: Roles and responsibilities
Part 311	Guidance for the application of product data quality for shape (PDQ-S)

A list of all the parts in the IS/ISO 8000 series can be found on the BIS and ISO websites.

This document specifies a procedure for data profiling to generate the foundation for performing data quality assessment. This profiling is applicable to data sets that are either originally in a structure of tables and columns or are the output from a transformation to create such a structure.

(*Continued on third cover*)

Contents

Page

Introduction

Digital data delivers value by enhancing all aspects of organizational performance including:

— operational effectiveness and efficiency;

— safety;

— reputation with customers and the wider public;

— compliance with statutory regulations;

— consumer costs, revenues and stock prices.

The influence on performance originates from data being the formalized representation of information; this information enables organizations to make reliable decisions. This decision making can be performed by human beings directly and also by automated data processing including artificial intelligence systems.

Through widespread adoption of digital computing and associated communication technologies, organizations become dependent on digital data. This dependency amplifies the negative consequences of lack of quality in this data. These consequences are the decrease of organizational performance.

The biggest impact of digital data comes from the data having a structure that reflects the nature of the subject matter and from the data also being computer processable (machine readable) rather than just being for a person to read and understand.

The content of ISO 9000 explains that quality is not an abstract concept of absolute perfection. Quality is actually the conformance of characteristics to requirements and, thus, any item of data can be of high quality for one use but not for another use that has differing requirements.

EXAMPLE 1 When storing start times for meetings, a calendar application requires less precision than a control system would for storing the times at which to activate a propulsion unit during a spaceflight.

The nature of digital data is fundamental to establishing requirements that are relevant to the specific decisions that are made by each organization.

EXAMPLE 2 ISO/TS 8000-1 identifies that data has syntactic (format), semantic (meaning) and pragmatic (usefulness) characteristics.

To support the delivery of high-quality data, the ISO 8000 series addresses:

— data governance, data quality management and maturity assessment;

EXAMPLE 3 ISO 8000-61 specifies a process reference model for data quality management.

— creating and applying requirements for data and information;

EXAMPLE 4 ISO 8000-110 specifies how to exchange characteristic data that is master data.

— monitoring and measuring data and information quality;

EXAMPLE 5 ISO 8000-8 specifies approaches to measuring data and information quality.

— improving data and, consequently, information quality;

EXAMPLE 6 This document specifies an approach to data profiling, which identifies opportunities to improve data quality.

— issues that are specific to the type of content in a data set.

EXAMPLE 7 ISO/TS 8000-311 specifies how to address quality considerations for product shape data.

Data quality management covers all aspects of data processing, including creating, collecting, storing, maintaining, transferring, exploiting and presenting data to deliver information.

Effective data quality management is systemic and systematic, requiring an understanding of the root causes of data quality issues. This understanding is the basis for not just correcting existing nonconformities but also implementing solutions that prevent future reoccurrence of those nonconformities.

EXAMPLE 8 If a data set includes dates in multiple formats including “yyyy-mm-dd”, “mm-dd-yy” and “dd-mm-yy”, then data cleansing can correct the consistency of the values. However, such cleansing requires additional information to resolve ambiguous entries (e.g. “04-05-20”) and cannot address any process issues and people issues, including training, that have caused the inconsistency.

As a contribution to this overall capability of the ISO 8000 series, this document specifies an approach to data profiling, which involves applying analysis techniques to data in actual use. This analysis generates a profile consisting of the structure, columns and relationships of the data. The profile provides the basis for identifying opportunities to improve data quality by establishing new explicit rules for the data. The approach also typically produces greater effect from repeated application to uncover issues progressively.

Organizations can use this document on its own or in conjunction with other parts of the ISO 8000 series.

This document supports activities that affect:

— one or more information systems;

— data flows within the organization and with external organizations;

— any phase of the data life cycle.

By implementing parts of the ISO 8000 series, an organization achieves the following benefits:

— establishing reliable foundations for digital transformation;

— recognizing how data in digital form has become a fundamental asset class that organizations rely on to deliver value;

— securing evidence-based trustworthiness of data and information for all stakeholders;

— creating portable data that protects against the loss of intellectual property and that is reusable across the organization and applications;

— achieving traceability of data back to original sources;

— ensuring all stakeholders work with common understanding of explicit data requirements.

ISO/TS 8000-1 provides a detailed explanation of the structure and scope of the ISO 8000 series.

Annex A contains an identifier that unambiguously identifies this document in an open information system.

Indian Standard

DATA QUALITY

PART 81 DATA QUALITY ASSESSMENT — PROFILING

1 Scope

This document specifies a procedure for data profiling to generate the foundation for performing data quality assessment. This profiling is applicable to data sets that are either originally in a structure of tables and columns or are the output from a transformation to create such a structure.

NOTE 1 Data profiling is applicable to all types of database technology.

The following are within the scope of this document:

— performing structure analysis to determine data element concepts;

— performing column analysis to identify relevant data elements, including statistics about a data set;

— performing relationship analysis to identify dependencies in a data set.

The following are outside the scope of this document:

— methods for extracting and sampling data to be profiled from a data set;

— deriving data rules;

— measuring the extent of nonconformities in a data set.

NOTE 2 ISO 8000-8 specifies approaches to measuring data and information quality.

This document can be used in conjunction with, or independently of, quality management systems standards.

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 8000-2, *Data quality — Part 2: Vocabulary*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 8000-2 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/

4 Data profiling

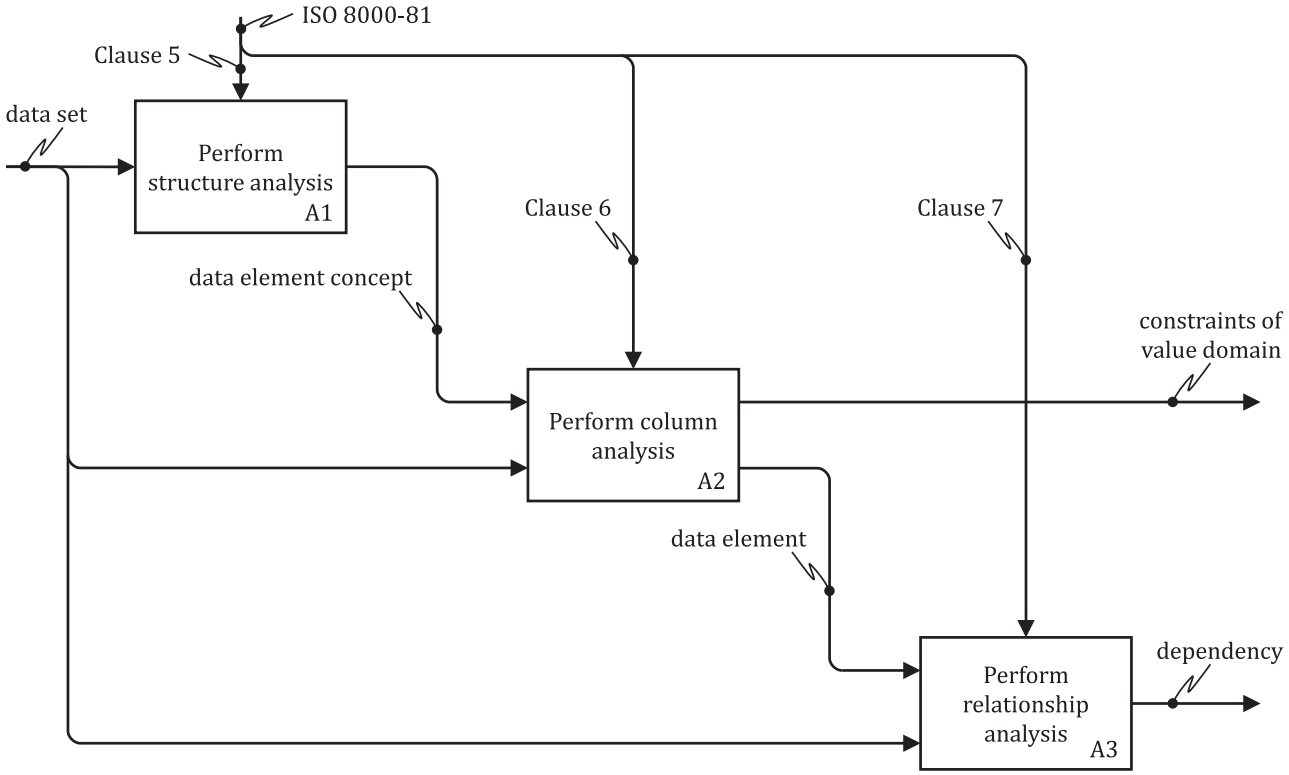
The purpose of data profiling is to characterize the structure, columns and relationships of a data set. This characterization is a data profile that serves as the basis on which an organization can improve data quality issues. The improvement can include creation of rules to enforce appropriate requirements on the data.

Data profiling consists of the following processes (see Figure 1):

— perform structure analysis (see Clause 5);

— perform column analysis (see Clause 6);

— perform relationship analysis (see Clause 7).



NOTE See ISO/IEC/IEEE 31320-1 for details on the notation used in this diagram.

Figure 1 — Perform data profiling

5 Structure analysis

5.1 Inputs

The input to structure analysis is a data set that consists of data values in one or more columns and, optionally, supporting information such as the name and description of each column.

5.2 Scope of activities

Structure analysis consists of:

— extracting the conceptual domain from the data values and any supporting information;

— determining the data element concept for use in column analysis (see Clause 6).

5.3 Outputs

The output from structure analysis is a data element concept.

6 Column analysis

6.1 Inputs

The inputs to column analysis are a data set and a corresponding data element concept from structure analysis (see Clause 5).

6.2 Scope of activities

Column analysis consists of:

— extracting data elements from the data element concept;

— comparing the data elements with the values in the data set;

— determining the value domain.

NOTE The methods for extracting data elements include discovery, assertion testing and visual inspection. These methods can be supported by automated tools.

6.3 Outputs

The output from column analysis is a list of constraints of value domain. These constraints include the following (see Annex B for more details):

— cardinalities: count of rows, range of values, nulls, count of distinct values and uniqueness;

— storage: data type, length of values and decimals;

— valid values: discrete value list, permissible range, skip-over rules, pattern and domain.

7 Relationship analysis

7.1 Inputs

The inputs for relationship analysis are a data set and the corresponding data elements from column analysis (see Clause 6).

NOTE Relationship analysis extracts relationships between columns within not only a single table but also multiple tables.

7.2 Scope of activities

Relationship analysis consists of:

— comparing the extracted data elements with any supporting information in the data set;

— determining dependency.

NOTE When performing relationship analysis, a key requirement is to understand the correspondence between the data structure (tables and columns) and items in the real world. This understanding arises from data profiling practitioners collaborating with experts who work with the core processes of the organization. These experts are familiar with the details of the items represented by the data.

7.3 Outputs

The output from relationship analysis is a list of dependencies, which include the following (see Annex C for more details):

— column dependencies: primary key, foreign key, functional dependency and derived column;

— synonyms: primary/foreign key synonym, redundant data synonym and domain synonym.

Annex A
(informative)

Document identification

To provide for unambiguous identification of an information object in an open system, the following object identifier is assigned to this document:

```
{ iso standard 8000 part(81) version(1) }
```

The meaning of this value is defined in ISO/IEC 8824-1 and is described in ISO 10303-1.

Annex B
(informative)

Constraints of value domain

The following constraints of value domain apply to sets of digital data.

— Cardinalities capture the overall range of values in a column (see Table B.1). This range establishes a basis on which to identify values that are potentially invalid because they are not consistent with the rest of the values in the column.

— Storage is a characterization of the fundamental rules for the syntax of values in a column (see Table B.2). These rules can be imposed by appropriate automated functionality of an information system, although often in practice such functionality is missing.

— Valid values are specific limits on which values are allowable in a column (see Table B.3). These limits can be more precise when the subject matter of the column is narrower.

EXAMPLE In general, a column *temperature* contains a more diverse range of values than a column *temperature in degrees Celsius* because the former can also include values in degrees Fahrenheit.

Table B.1 — Constraints of value domain: Cardinalities

Constraint	Description	Role	Example
Count of rows	The total number of individual values in a column, including nulls and duplicates.	Establishes the denominator for any calculations about the ratio of individual values to the total population.	A result expressed as a single integer (e.g. 3177).
Range of values	The statistical characterization of the population of values in a column.	Establishes a baseline understanding of the data currently in a column.	Results for the minimum, maximum, median and mean of the values in a column.
Nulls	The number of values that contain no data (i.e. are blank or some other similar representation of the absence of data).	Helps to discover whether the column has the attribute of being mandatory, optional or conditional.	A result expressed as an absolute number (e.g. 2769) of values that are null. A result expressed as a percentage (0 % to 100 %) of values that are null.
Count of distinct values	The size of the set of values after removing all but one of each duplicate value.	Helps to discover the domain of a column.	When the complete set of values in a column consists of “100”, “100”, “200”, “200” and “300”, then the result is 3.
Uniqueness	The degree to which each value in a column is unique.	Helps to discover columns that contain primary keys.	A result expressed as a percentage (0 % to 100 %) of values that are unique.

Table B.2 — Constraints of value domain: Storage

Constraint	Description	Role	Example
Data type	The nature of the value.	Enforces all values to the same type.	The column constraints CHARACTER, INTEGER, DECIMAL, DATE, TIME, TIMESTAMP, BINARY and DOUBLEBYTE.
Length of values	The number of digits or characters that may form in a value.	Limits the length (either as an absolute or as a maximum).	The column constraints VARIABLE, FIXED 5 and NUMERIC 5.
Decimals	The maximum number of decimal places for numeric values.	Enforces a precision that is appropriate to the use of the data.	The column constraint DECIMAL 2.

Table B.3 — Constraints of value domain: Valid values

Constraint	Description	Role	Example
Discrete value list	A list of a small number of specific values.	Avoids users entering levels of detail that are inappropriate for the use of the data.	For an information system recording missing luggage items for an airline, only listing simple colours such as “black”, “blue” and “brown” for the column *colour of missing luggage*.
Permissible range	Defines valid values to lie between a minimum and a maximum.	Limits values to a range that reflects the nature of the item described by the data.	For an information system recording weather conditions on the Earth, “-100” to “+100” for the range of the column *outside air temperature (degree Celsius)*.
Skip-over rules	Excludes specific values.	Limits the range of values in a column.	For an information system supporting a courier company delivering parcels on working days, excluding weekends and holidays from the column *expected delivery date*.
Pattern	Defines a syntax for a value in terms of valid ranges of characters in individual positions within the value.	Without guaranteeing the existence of the value, prevents the user from entering a value that is fundamentally incorrect for the column.	For an information system recording contact details for persons, only accepting values with the pattern <name> “@” <fully qualified domain name> in the column *e-mail address* (i.e. additional validation is necessary to check whether each e-mail address actually exists).
Domain	Set of unique, distinct permissible values.	Limits values to those appropriate to the nature of the item identified by the data.	Permissible values “male” and “female” for the column *sex*. Permissible values for the column *credit card type* corresponding to the companies providing credit card processing services.

Annex C
(informative)

Dependency

A dependency exists between two or more columns in a data set. There are two key categories of dependency:

— column dependencies (see Table C.1), where the relationship between columns is supporting the coherence of the structure of the data set;

— synonyms (see Table C.2), where the columns represent the same item in the real world.

Table C.1 — Column dependencies

Dependency	Description	Role	Example
Primary key	One or more columns that uniquely define each row of a table.	Identifies each row of a table, enabling relationships from other tables in the data set.	SOCIAL_SECURITY_NUMBER PERSON_ID
Foreign key	One or more columns in a dependent table that identify a row in a parent table.	Establishes a parent/dependency relationship between two tables.	In a dependent table *departments*, the column DEPT_MANAGER_ID is the foreign key relating to the primary key PERSON_ID in the parent table *personnel*.
Functional dependency	A column has a functional dependency on one or more other columns in the same table if the value is determined by the values in one or more other columns.	Indicates that a column value is not independent of other columns in a table.	The values in the columns TEMP_DEG_CELSIUS and TEMP_DEG_FAHRENHEIT are dependent through a formula for temperature conversion.
Derived column	A column is the output of a function that takes values in one or more other columns as inputs.	Provides the basis on which to avoid storage of redundant data and instead generate values by algorithm executed by the information system.	A user interface takes temperature in degrees Celsius as input, stores that value in a column TEMP_DEG_CELSIUS and generates a value for TEMP_DEG_FAHRENHEIT, which is the derived column.

Table C.2 — Synonyms

Dependency	Description	Role	Example
Primary/foreign key synonym	When a column is a primary key and another column is a foreign key, the two columns are primary/foreign key synonyms.	Gives a parent/dependency relationship between two tables.	In a dependent table *departments*, the column DEPT_MANAGER_ID is the foreign key relating to the primary key PERSON_ID in the parent table *personnel*. These two columns are primary/foreign key synonyms.
Redundant data synonym	If one column is a synonym with another column in a different table, and when the column is dependent on a key as well as is dependent on a corresponding foreign key in the different table, the two columns are in the relationship of redundant data synonym.	Provides the basis on which to re-configure the information system to display values based on data lookup rather than duplicated storage of those values.	By being able to view the description of a product in the record of an order, the person processing the order can more readily confirm the details of the order. This description is, however, a common value that the table *products* stores (see Figure C.1).
Domain synonym	When two columns are domain synonyms, the domain of one column is the same as that of the other column.	These columns are referring to the same items in the real world and, thus, offer the opportunity to establish the same mechanism by which to define and control the values in each column.	The tables *customers* and *employees* both include a column CITY (see Figure C.2).

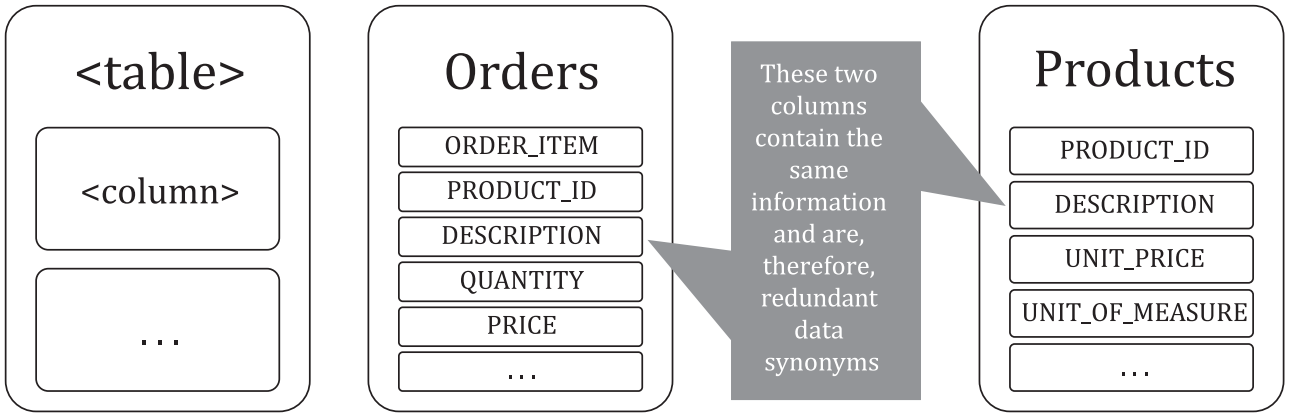


Figure C.1 — Example of a redundant data synonym

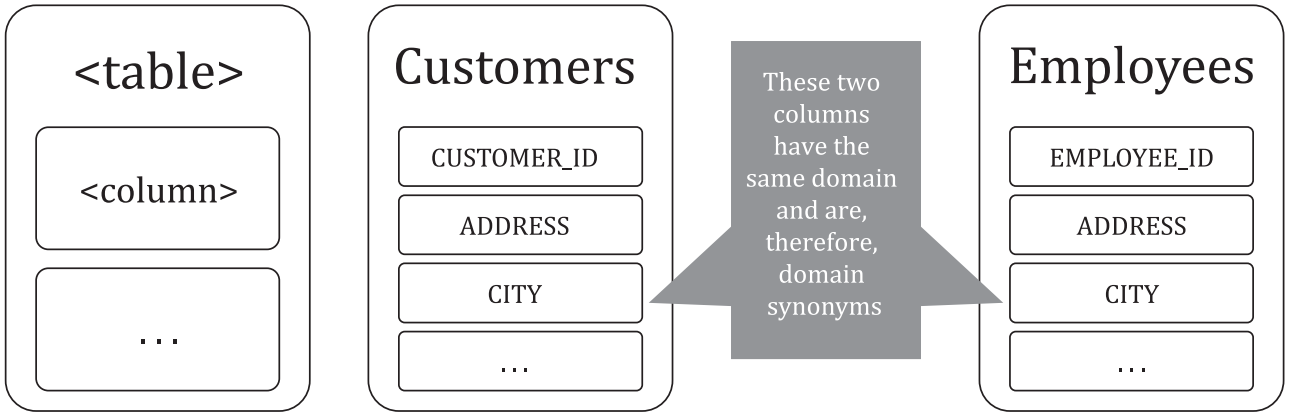


Figure C.2 — Example of a domain synonym

Bibliography

[1] ISO/TS 8000-1, *Data quality — Part 1: Overview*

[2] ISO 8000-8, *Data quality — Part 8: Information and data quality: Concepts and measuring*

[3] ISO 8000-61, *Data quality — Part 61: Data quality management: Process reference model*

[4] ISO 8000-110, *Data quality — Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification*

[5] ISO/TS 8000-311, *Data quality — Part 311: Guidance for the application of product data quality for shape (PDQ-S)*

[6] ISO 9000, *Quality management systems — Fundamentals and vocabulary*

[7] ISO/IEC/IEEE 31320-1, *Information technology — Modeling Languages — Part 1: Syntax and Semantics for IDEF0*

(*Continured from second cover*)

The text of ISO standard has been approved as suitable for publication as an Indian Standard without deviations. Certain conventions are, however, not identical to those used in Indian Standards. Attention is particularly drawn to the following:

a) Wherever the words 'International Standard' appear referring to this standard, they should be read as 'Indian Standard'; and
b) Comma (,) has been used as a decimal marker while in Indian Standards, the current practice is to use a point (.) as the decimal marker.

In this adopted standard, reference appears to the following International Standard for which Indian Standard also exists. The corresponding Indian Standard which is to be substituted in its place is given below along with its degree of equivalence for the edition indicated.

International Standard	*Corresponding Indian Standard*	*Degree of Equivalence*
ISO 8000-2 Data quality — Part 2: Vocabulary	IS/ISO 8000-2 : 2022 Data quality — Part 2 Vocabulary	Identical

For the purpose of deciding whether a particular requirement of this standard is complied with, the final value, observed or calculated expressing the result of a test or analysis, shall be rounded off in accordance with IS 2 : 2022 'Rules for rounding off numerical values (*second revision*).' The number of significant places retained in the rounded off value should be the same as that of the specified value in this standard.

Bureau of Indian Standards

BIS is a statutory institution established under the *Bureau of Indian Standards Act*, 2016 to promote harmonious development of the activities of standardization, marking and quality certification of goods and attending to connected matters in the country.

Copyright

Review of Indian Standards

Amendments are issued to standards as the need arises on the basis of comments. Standards are also reviewed periodically; a standard along with amendments is reaffirmed when such review indicates that no changes are needed; if the review indicates that changes are needed, it is taken up for revision. Users of Indian Standards should ascertain that they are in possession of the latest amendments or edition by referring to the website-www.bis.gov.in or www.standardsbis.in.

This Indian Standard has been developed from Doc No.: PGD 18 (22178).

Amendments Issued Since Publication

Amend No.	Date of Issue	Text Affected

BUREAU OF INDIAN STANDARDS

Headquarters:

Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002
Telephones: 2323 0131, 2323 3375, 2323 9402 *Website*: www.bis.gov.in

Regional Offices:		*Telephones*
Central	: 601/A, Konnectus Tower -1, 6th Floor, DMRC Building, Bhavbhuti Marg, New Delhi 110002	2323 7617
Eastern	: 8th Floor, Plot No 7/7 & 7/8, CP Block, Sector V, Salt Lake, Kolkata, West Bengal 700091	2367 0012 2320 9474
Northern	: Plot No. 4-A, Sector 27-B, Madhya Marg, Chandigarh 160019	265 9930
Southern	: C.I.T. Campus, IV Cross Road, Taramani, Chennai 600113	2254 1442 2254 1216
Western	: Plot No. E-9, Road No.-8, MIDC, Andheri (East), Mumbai 400093	2821 8093

Branches : AHMEDABAD. BENGALURU. BHOPAL. BHUBANESHWAR. CHANDIGARH. CHENNAI. COIMBATORE. DEHRADUN. DELHI. FARIDABAD. GHAZIABAD. GUWAHATI. HIMACHAL PRADESH. HUBLI. HYDERABAD. JAIPUR. JAMMU & KASHMIR. JAMSHEDPUR. KOCHI. KOLKATA. LUCKNOW. MADURAI. MUMBAI. NAGPUR. NOIDA. PANIPAT. PATNA. PUNE. RAIPUR. RAJKOT. SURAT. VISAKHAPATNAM.

Published by BIS, New Delhi