TECHNICAL SPECIFICATION

ISO/TS 8000-82

First edition
2022-06

Data quality —

Part 82:
Data quality assessment: Creating data rules

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO documents 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).

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

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.

This document was prepared by Technical Committee ISO/TC 184, *Automation systems and integration*, Subcommittee SC 4, *Industrial data*.

A list of all parts in the ISO 8000 series can be found on the ISO website.

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.

Introduction

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

— operational effectiveness and efficiency;

— safety and security;

— reputation with customers and the wider public;

— compliance with statutory regulations;

— innovation;

— consumer costs, revenues and stock prices.

In addition, many organizations are now addressing these considerations with reference to the United Nations Sustainable Development Goals[1].

The influence on performance originates from data being the formalized representation of information[2]. This information enables organizations to make reliable decisions. Such 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 these data. These consequences are the decrease of organizational performance.

The biggest impact of digital data comes from two key factors:

— the data having a structure that reflects the nature of the subject matter;

EXAMPLE 1   A research scientist writes a report using a software application for word processing. This report includes a table that uses a clear, logical layout to show results from an experiment. These results indicate how material properties vary with temperature. The report is read by a designer, who uses the results to create a product that works in a range of different operating temperatures.

— the data being computer processable (machine readable) rather than just being for a person to read and understand.

EXAMPLE 2   A research scientist uses a database system to store the results of experiments on a material. This system controls the format of different values in the data set. The system generates an output file of digital data. This file is processed by a software application for engineering analysis. The application determines the optimum geometry when using the material to make a product.

ISO 9000 explains that quality is not an abstract concept of absolute perfection. Quality is actually the conformance of characteristics to requirements. This actuality means that any item of data can be of high quality for one purpose but not for a different purpose. The quality is different because the requirements are different between the two purposes.

EXAMPLE 3   Time data are processed by calendar applications and also by control systems for propulsion units on spacecraft. These data include start times for meetings in a calendar application and activation times in a control system. These start times require less precision than the activation times.

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

EXAMPLE 4   ISO 8000-8 identifies that data have syntactic (format), semantic (meaning) and pragmatic (usefulness) characteristics.

---

1)   https://sdgs.un.org/goals

2)   ISO 8000-2 defines information as "knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning".

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

— data governance, data quality management and maturity assessment;

EXAMPLE 5   ISO 8000-61 specifies a process reference model for data quality management.

— creating and applying requirements for data and information;

EXAMPLE 6   ISO 8000-110 specifies how to exchange characteristic data that are master data.

— monitoring and measuring information and data quality;

EXAMPLE 7   ISO 8000-8 specifies approaches to measuring information and data quality.

— improving data and, consequently, information quality;

EXAMPLE 8   ISO/TS 8000-81 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 9   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 10   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. Such cleansing requires additional information, however, to resolve ambiguous entries (e.g. "04-05-20"). The cleansing also 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 the characteristics of data rules that can support data quality assessment.

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 to improve organizational performance, an organization achieves the following benefits:

— objective validation of the foundations for digital transformation of the organization;

— a sustainable basis for data in digital form becoming a fundamental asset class the organization relies on to deliver value;

— securing evidence-based trust from other parties (including supply chain partners and regulators) about the repeatability and reliability of data and information processing in the organization;

— portability of data with resulting protection against loss of intellectual property and reusability across the organization and applications;

— effective and efficient interoperability between all parties in a supply chain to achieve traceability of data back to original sources;

— readiness to acquire or supply services where the other party expects to work with common understanding of explicit data requirements.

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

ISO 8000-2[3] specifies the single, common vocabulary for the ISO 8000 series. This vocabulary is a foundation for understanding the overall subject matter of data quality. ISO 8000-2 presents the vocabulary structured by a series of topic areas (for example, terms relating to quality and terms relating to data and information).

ISO has identified ISO 8000-1, ISO 8000-2 and ISO 8000-8 as horizontal deliverables[4].

Annex A contains an identifier that conforms to ISO/IEC 8824-1. The identifier unambiguously identifies this document in an open information system.

---

[3] The content is available on the ISO Online Browsing Platform. https://www.iso.org/obp

[4] Deliverable dealing with a subject relevant to a number of committees or sectors or of crucial importance to ensure coherence across standardization deliverables.

vii

Data quality —

Part 82:
Data quality assessment: Creating data rules

1   Scope

This document describes how data rules apply to various types of data. Such rules exist to sustain the integrity and reliability of data by capturing requirements into a form that can be processed by databases and other information systems.

The following are within the scope of this document:

— fundamental concepts of data rules;

— key characteristics of data rules for common types of data, where these types are identifier, currency value, quantity, date or time, rate, free-text entry, code and key;

— how data profiling contributes to formulating effective data rules.

The following is outside the scope of this document:

— specific rules for specific sets of data.

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

EXAMPLE 1   ISO 9001 specifies requirements for quality management systems.

This document can also be used in conjunction with or independently of standards for more detailed definitions of data types.

EXAMPLE 2   ISO/IEC 11404 specifies the nomenclature and shared semantics for a collection of datatypes commonly occurring in programming languages and software interfaces.

EXAMPLE 3   IEC 61360-1 specifies principles for the definition of the properties and associated attributes and explains the methods for representing verbally defined concepts.

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*

ISO/TS 8000-81, *Data quality — Part 81: Data quality assessment: Profiling*

3   Terms and definitions

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

1

— IEC Electropedia: available at https://www.electropedia.org/

4 Basic concepts for data rules

To gain sufficient evidence necessary to create or improve data rules for data, an organization shall perform data profiling in accordance with ISO/TS 8000-81. A data rule is the specification that applies to data values that are instances of a particular data attribute within a data set.

EXAMPLE 1   A data set includes information on five different people. The set includes the instance values "female", "female", "male", "female" and "male" for the data attribute "sex". A rule states that "female", "male" and "non-binary" are allowed values for the attribute.

Each data rule:

— captures the results of any design analysis to create a database in which to store the value;

EXAMPLE 2   A database designer decides whether each column in a database table is either mandatory or optional.

— reflects requirements corresponding to the purpose of the values;

EXAMPLE 3   To enable engineers to select the correct bolts for different locations within a complex mechanical system, separate data rules specify the required number of decimal places for a numeric value and the required unit of measurement that together represent the length of each bolt listed by a parts catalogue.

— enables organizations to assess quality of values in the database and to identify which values do not conform to the applicable requirements;

NOTE 1   Such assessment is an important part of proactive data quality management.

— can apply to any aspect of data processing that occurs during the data lifecycle;

NOTE 2   The lifecycle covers creation to destruction of a data value.

— covers various different characteristics of the data value;

EXAMPLE 4   These characteristics can include whether the value is mandatory, the maximum and minimum allowable magnitude of the value and restrictions on the formatting of the value (such as consisting of exactly four characters, each of which must be in the range "A" to "Z").

— can originate from overall policy controlling data management, the specific design analysis task or the specific functionality of the database system.

NOTE 3   The existence of multiple data rules requires an integrated approach to managing the rules as a single coherent set.

5 Data rules

5.1 Overview

This document specifies the key characteristics of data rules for the following types of data:

— identifier (see 5.2);

— currency value (see 5.3);

— quantity (see 5.4);

— date or time (see 5.5);

— rate (see 5.6);

— free-text entry (see 5.7);

— code (see 5.8);

— key (see 5.9).

NOTE   These are only common types of data, where additional types are useful for different uses of data. These additional types are specified by other standards including ISO/IEC 11404 and IEC 61360-1.

5.2 Data rules for identifiers

The key characteristics of a data rule for an identifier are:

— the rule can specify which characters are allowed to form the identifier;

EXAMPLE 1   A rule specifies the allowable characters to be from the Latin alphabet (lower and upper case), the Arabic numerals (i.e. "0" to "9") and some characters that serve as separators ("#" and "-").

— the identifier can identify a specific collection of data values according to either a scheme specific to the system storing the data set or a scheme that exists independently of any one system;

— the identifier can uniquely identify the subject matter of the collection of data values;

EXAMPLE 2   The subject matter is an object, person or organization.

— the scheme for the identifier can be specific to the organization, be the result of national collaboration or be the result of international collaboration;

— the rule ensures the data value representing the identifier has a format that conforms to the requirements of the identifier scheme.

Data profiling can help an organization to generate data rules for identifiers when the organization does not know the scheme that applies to the identifier. The profiling can identify those identifier values that appear anomalous with respect to the general pattern of the majority of the values in a data set. These anomalies can be the focus for further investigation, avoiding the organization from wasting scarce resources on assessing the quality of excessive amounts of data.

EXAMPLE 3   Identifiers include International Standard Book Number, resident registration number, business registration number, postal code, customer number, account number, Uniform Resource Locator, email address and Internet Protocol address.

EXAMPLE 4   ISO 8000-115 specifies rules for identifiers of master data when exchanging those data.

5.3 Data rules for currency values

The key characteristics of a data rule for a currency value are:

— the value indicates the amount and the currency of the amount;

— the rule ensures the value is within a range that is consistent with the meaning of the data attribute.

EXAMPLE 1   A rule allows the price of an item only to be positive, but another allows the balance of a bank account to be negative.

Data profiling can help an organization to generate data rules for currency values by considering:

— individual values after analysing the minimum value, maximum value, mean, variance and standard deviation in a data set;

— related values after checking for consistency with the underlying calculation that supposedly has generated a dependent value.

EXAMPLE 2   For the balance of a bank account on two different dates, is the change in the balance consistent with all the credit and debit entries for the account between those dates.

EXAMPLE 3   Currency values include tax, price, unit price, cost, charges, account balances and inventory value totals.

5.4   Data rules for quantities

A quantity is a property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed by means of a number and a unit of measurement, measurement procedure, reference material or combination of such.

NOTE   ISO 80000-1 provides further details on quantities and units of measurement.

The key characteristics of a data rule for a quantity are:

— the value is a number that is appropriate for the associated unit of measurement, measurement procedure, reference material or combination of such;

EXAMPLE 1   A quantity is a number that is a count of items and has an associated unit of measurement of "each".

— the rule ensures the value is in a valid range.

Data profiling can help an organization to generate data rules for quantities by analysing the minimum value, maximum value, mean, variance and standard deviation in a data set.

EXAMPLE 2   Quantities include number of events, number of sheets, number of items, number of times, distance, scale, length, weight, speed, area and temperature.

5.5   Data rules for dates and times

The key characteristics of a data rule for a date or time are:

— the value represents an instant or a time interval;

EXAMPLE 1   The ISO 8601 series specifies the expression "1968-07-17" as representing the 24-hour period of time at the point in time of day 17 of month of July in the year 1968 according to the Gregorian calendar.

EXAMPLE 2   The value indicates the occurrence of some event, such as date of receipt, date of registration, month of settlement, and time of transmission.

— the rule establishes whether the value is a date, time or date and time;

EXAMPLE 3   The expression "1968-07-17" is a date; this expression is different than the expression "1968-07-17T00:00:00Z", which is a date and time. The rule ensures applications do not convert a date into a date and time, thus adding inappropriate information to the value.

— the rule ensures the structure of the value is appropriate to represent the date or time and ensures each element of the structure is consistent with the meaning;

EXAMPLE 4   In the representation "YYYY-MM-DD" specified by the ISO 8601 series, the element "MM" contains one of the values in the set {"01", "02" … "12"}.

EXAMPLE 5   In order to prevent ambiguity of date and time expressions, the ISO 8601 series specifies how to include an indication of time zone in those expressions.

— the rule provides the basis for controlling the manual input of dates or times.

Data profiling can help an organization to generate data rules for dates by checking for variation in formats between different information systems across the organization. Such variation causes a risk that values are changed by exchange between systems and that analysis does not draw the correct inference from the data. The organization can remove this risk by implementing standard date formats for all systems.

EXAMPLE 6   Dates include year-month, year, year-month-day, hour, minute, second, day, half year and quarter.

5.6 Data rules for rates

The key characteristics of a data rule for a rate are:

— the value can either be a derivation from other numeric data or be a master value that determines results when processing other data within an organization;

EXAMPLE 1 Derivation can result in values such as growth rates, return rates and fluctuation rates. Master values include interest rates for either saving or borrowing money.

— the many varieties of rate prevent there being a single, typical rule for rates and, instead, the details of each rule depend on the purpose of the rate;

— the rule can prevent significant consequences from monetary losses when poor quality of a rate affects financial or accounting activities within an organization.

Data profiling can help an organization to generate data rules for rates by analysing the minimum value, maximum value, mean, variance and standard deviation in a data set.

EXAMPLE 2 Rates include interest rate, ratio, exchange rate and percentage.

5.7 Data rules for free text entries

For a data rule applying to a free text entry, the key characteristics are:

— the value consists of a text in natural language using a set of characters that support the applicable writing system;

— the text can vary from relatively short content to a long and comprehensive description and can support many different purposes, although people are likely to be the primary audience for the content in the first instance;

— the rule can make use of predefined lists of allowable content to control some issues but, in general is not able to prevent more complex issues such as the text being ambiguous;

EXAMPLE 1 The most general type of predefined list is the dictionary for the applicable natural language. Such a dictionary can also be supported by some patterns to indicate the different forms in which a verb can appear in a text.

— the rule can enforce minimum and maximum length of the value in order to reduce the threat that the content is not sufficient, concise and relevant.

Data profiling can help an organization to generate data rules for free text entries by providing evidence as to whether the value is to be mandatory or, if the content is to provide supplementary information, optional. The profiling can also identify where users have used content without meaning to circumvent the value being mandatory.

EXAMPLE 2 When a value is optional then the content can be NULL, empty or blank.

EXAMPLE 3 Content without meaning can include space characters or just punctuation such as the "." character.

EXAMPLE 4 Free text entries include comments, descriptions, diagnostic information and summaries.

5.8 Data rules for codes

The key characteristics of a data rule for a code are:

— the value is a simplified, consistent representation of an underlying concept;

EXAMPLE 1 Codes include customer rating code, department code, product code and area code.

— typically, the value is part of a list that the organization will not expect to continue to grow without limit;

EXAMPLE 2   If an organization allows users to enter classification values as free-text entries then this will result in proliferation, for instance from the use of synonyms such as "temporary export" and "non-permanent export" in completing a customs form. If the organization defines codes, then a user requests a new code if the existing codes do not include an appropriate value.

— a binary classifier is a common type of code;

EXAMPLE 3   Typical binary classifiers include {"Yes", "No"}, {"Pass", "Fail"} and {1, 0}.

— by using codes, organizations increase the efficiency and accuracy of manual data entry;

— organizations can support users to understand codes by providing short names, long names, descriptions and all of these items in multiple languages if the code supports operations in more than one country.

EXAMPLE 4   The codes "M" and "F" can represent "male" and "female" respectively, but the organization can also provide translations into French ("masculin" and "féminin" respectively).

Data profiling can help an organization to generate data rules for codes by detecting patterns in the occurrence of values in a data set and identifying how an unknown value could be a corruption of a valid code. This profiling enables the organization to reduce the risk of unknown codes affecting the ability of the organization to combine different data sets and to make accurate decisions.

EXAMPLE 5   Profiling detects an unknown value "PO1" (using the upper-case letter "O") but the valid codes include the value "P01", using the digit "0" (zero).

EXAMPLE 6   Codes include individual code and integrated code.

EXAMPLE 7   Codes that are binary classifiers include status and category and apply to attributes such as smoking status, drinking status, home owner status and marketing contact status.

5.9   Data rules for keys

The key characteristics of a data rule for a key are:

— the main feature of each value is the role that the value plays in connecting two or more different parts of an overall data set;

EXAMPLE   A foreign key in a database table points at the primary key in another table.

— the value is either a primary key to identify data, a unique key to prevent duplication of data or a foreign key to represent a relationship between data items;

— the rule can operate either at the physical level to preserve integrity of database systems or at the logical level to address the implications of using the data across multiple systems.

Data profiling can help an organization to generate data rules for keys by checking for repetition of values (to identify discrepancies against a requirement of uniqueness) and for mismatch between sets of values (to identify where an intended relationship is incomplete because a key value only exists at one end of that relationship).