TECHNICAL SPECIFICATION

ISO/TS 8000-81

First edition
2021-05

Data quality —

Part 81:
Data quality assessment: Profiling

Reference number
ISO/TS 8000-81:2021(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) 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 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.

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

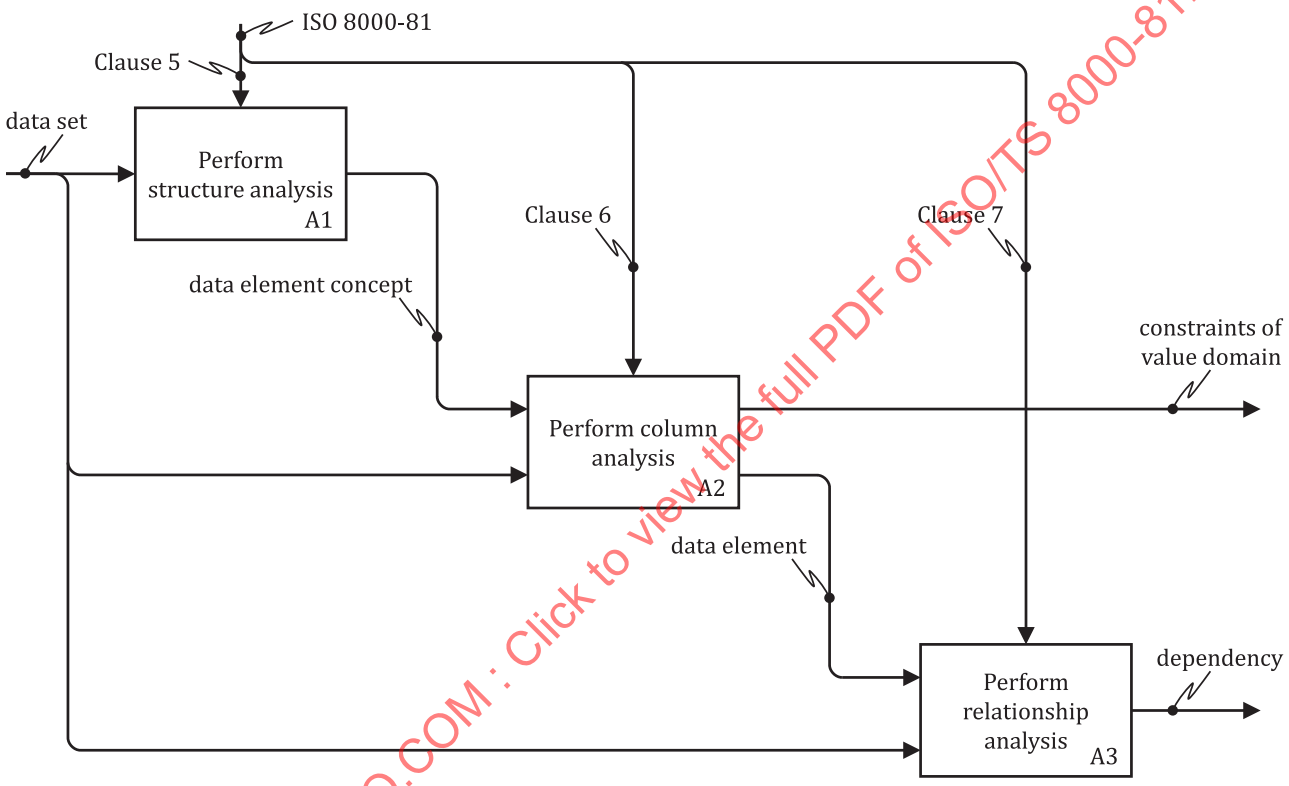
The purpose of data profiling is to characterize the structure, columns and relationships of a data set. This characterization is a data profile that serves as the basis on which an organization can improve data quality issues. The improvement can include creation of rules to enforce appropriate requirements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