

---

---

**Data quality —**  
Part 81:  
**Data quality assessment: Profiling**

**iTeh STANDARD PREVIEW**  
**(standards.iteh.ai)**

[ISO/PRF TS 8000-81](https://standards.iteh.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81)

<https://standards.iteh.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81>

**PROOF / ÉPREUVE**

---

---



**iTeh STANDARD PREVIEW**  
**(standards.iteh.ai)**

[ISO/PRF TS 8000-81](https://standards.iteh.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81)

<https://standards.iteh.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81>



**COPYRIGHT PROTECTED DOCUMENT**

© ISO 2021

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](mailto:copyright@iso.org)  
Website: [www.iso.org](http://www.iso.org)

Published in Switzerland

# Contents

	Page
Foreword .....	iv
Introduction .....	v
<b>1 Scope</b> .....	<b>1</b>
<b>2 Normative references</b> .....	<b>1</b>
<b>3 Terms and definitions</b> .....	<b>1</b>
<b>4 Data profiling</b> .....	<b>2</b>
<b>5 Structure analysis</b> .....	<b>2</b>
5.1 Inputs .....	2
5.2 Scope of activities .....	2
5.3 Outputs .....	3
<b>6 Column analysis</b> .....	<b>3</b>
6.1 Inputs .....	3
6.2 Scope of activities .....	3
6.3 Outputs .....	3
<b>7 Relationship analysis</b> .....	<b>3</b>
7.1 Inputs .....	3
7.2 Scope of activities .....	3
7.3 Outputs .....	4
<b>Annex A (informative) Document identification</b> .....	<b>5</b>
<b>Annex B (informative) Constraints of value domain</b> .....	<b>6</b>
<b>Annex C (informative) Dependency</b> .....	<b>8</b>
<b>Bibliography</b> .....	<b>11</b>

iTech STANDARD PREVIEW  
(standards.itech.ai)

[ISO/PRF TS 8000-81](https://standards.itech.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81)

<https://standards.itech.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81>

## 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](http://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](http://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](http://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](http://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.

## ISO/TS 8000-81:2021(E)

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:

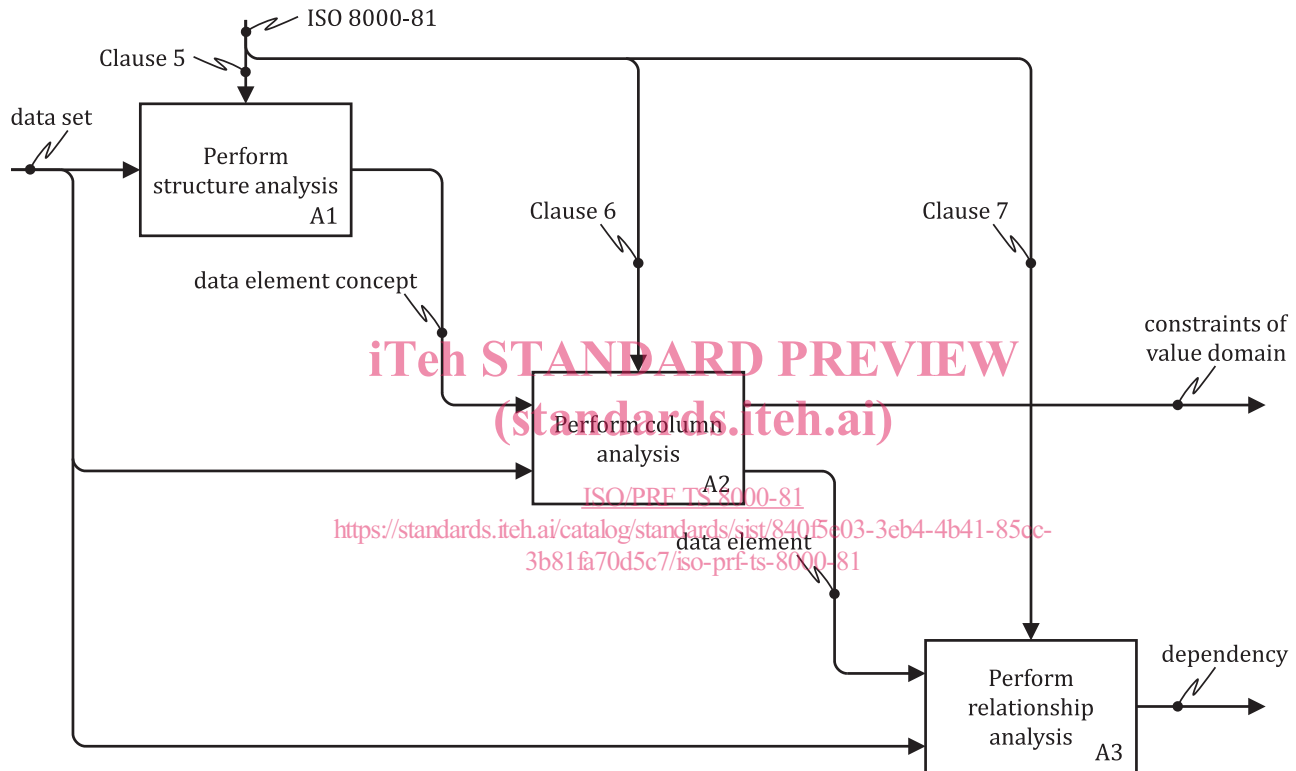
- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <http://www.electropedia.org/>

## 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.

## iTeh STANDARD PREVIEW (standards.iteh.ai)

[ISO/PRF TS 8000-81](#)

<https://standards.iteh.ai/catalog/standards/sist/840f5e03-3eb4-4b41-85cc-3b81fa70d5c7/iso-prf-ts-8000-81>