TECHNICAL REPORT

ISO/IEC TR 9789

First edition
1994-12-15

Information technology — Guidelines for the organization and representation of data elements for data interchange — Coding methods and principles

Technologies de l'information — Principes directeurs pour l'organisation et la représentation des éléments de données pour l'échange de données — Méthodes et principes de codage

Reference number
ISO/IEC TR 9789:1994(E)

Contents

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization.
National Bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC/JTC 1.

The main task of technical committees is to prepare International Standards. In exceptional circumstances a technical committee may propose the publication of a Technical report of one of the following types:

- type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

- type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility for an agreement on an International Standard;

- type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard; ("state of the art", for example).

Technical reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transferred into International Standard. Technical reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

ISO/IEC TR 9789, which is a Technical Report of type 3, was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology* Subcommittee 14, *Data element principles.*

Introduction

This Technical Report is a guide to develop and implement coded representations.
Coding covers the way and the form in which data are expressed.

The increased use of data processing and electronic data interchange heavily relies on accurate, reliable, controllable and verifiable data recorded in data bases.
In formal communication and storage data are expressed in symbols (usually digits or letters), arithmetic numbers and descriptions, which should have a fixed stable meaning for every one involved and thereby be suitable for purposes of processing and communication.

This Technical Report presents the objectives of coding, the characteristics, advantages and disadvantages of different coding methods, a survey of the features of codes and guidelines for the design of codes.

This page intentionally left blank

Information technology -- Guidelines for the organization and representation of data elements for data interchange -- Coding methods and principles

1 Scope

This Technical Report provides general guidance on the manner on which data can be expressed by codes.

It describes the objectives of coding, the characteristics, advantages and disadvantages of different coding methods, the features of codes and gives guidelines for the design of codes.

This Technical Report is not directed toward any specific application area nor dependent on any design method for application systems or data interchange.

2 References

2.1 General references

ISO/IEC 646:1991, *Information technology - ISO 7-bit coded character set for information interchange.*

ISO 2382-4:1987, *Information processing systems - Vocabulary - Part 04:Organization of data.*

ISO 2375:1985, *Data processing - Procedure for registration of escape sequences.*

ISO 7064:1983, *Data processing - Check character systems.*

ISO/IEC 11179-3:1994, *Information technology - Specification and standardization of data elements - Part 3: Basic attributes of data elements.*

2.2 Examples of applications of this Technical Report

ISO 9735:1988, *Electronic data interchange for administration, commerce and transport (EDIFACT) - Application level syntax.*

ISO 8601:1988, *Data elements and interchange formats - Information interchange - Representation of dates and times.*

ISO 3166:1993, *Codes for the representation of names of countries.*

3 Definitions

For the purpose of this Technical Report, the following definitions apply.

3.1 attribute: A characteristic of an object.

3.2 character set: A finite set of different characters that is complete for a given purpose.

Example: The international reference version of the character set of ISO 646.

3.3 code: A collection of rules that maps the elements of one set on to the elements of another set.

NOTES

1. The elements may be characters or character strings.
2. The first set is the *coded set* and the other set is the *code element set.*
3. An element of the code element set may be related to more than one element of the coded set but the reverse is not true.

3.4 code element: The result of applying a *code* to an element in a *coded set.*

Examples:

1 "CDG" as the representation of Paris Charles de Gaulle in the code for three-letter representation of airport names.

2 The seven binary digits representing the delete character in ISO 646.

3.5 code element set: The result of applying a *code* to all elements of a *coded set.*

Example: All the three-letter international representations of airport names.

3.6 code set: Synonym of **code element set.**

3.7 code value: Synonym of **code element.**

3.8 coded representation: Synonym of **code element.**

3.9 coded set: A set of elements which is mapped on to another set according to a *code.*

Example: A list of the names of airports which is mapped on to a corresponding set of three-letter abbreviations.

3.10 coding scheme: Synonym of **code.**

3.11 data code: See preferred term **code element.**

3.12 data element instance: An occurrence of a *data element type.*

3.13 data element type: A category of data which represents a concept and whose properties are expressed as a set of data element type attributes which permit it to support information interchange.

3.14 delimiter: One or more characters used to indicate the beginning or the end of a character string.

3.15 entity: Any concrete or abstract thing of interest, including associations among things.

3.16 field: A specified area on a data medium or in storage, used for a particular class of data elements.

3.17 identifier: One or more characters used to identify or name a data element type and possibly to indicate certain properties of that data element type.

3.18 key: An identifier within a set of data element types.

3.19 position: Any location in a *string* that may be occupied by an element and that is identified by a serial number.

3.20 string: A sequence of elements of the same nature, such as characters, considered as a whole.

3.21 table: An arrangement of data each item of which may be identified by means of arguments or keys.

4 Principles of coding

4.1 Information and coding

In daily life information is understood as facts of and propositions about all the concrete or abstract things of interest expressed by: data, messages and further particulars.
Information is necessary for the proper execution of any conceivable task be it in administration, commerce, transport, science, etc.
Accurate, objective and unambiguous information is a prerequisite in cases of computer based information systems and the data interchange between them.

In formal information systems data is expressed in symbols (usually digits or letters), arithmetic numbers and descriptions, which should have a fixed, stable meaning for every one involved and thereby be suitable for purposes of processing and communication.

Users, irrespective of their function or tasks, should be able to understand, interpret and handle their information correctly.
Information shared by different user groups or application systems has to have an agreed definition e.g. semantic meaning of a concept (connotation) and all instances of a concept (denotation) and an agreed representation.

Coding covers the way and the form in which data is expressed by codes.
It is necessary to make clear agreements on these representations.
An explanation of the representation forms and code elements is part of the specification of data.

4.2 Coding

By coding is understood the rule-based assignment of code elements to a named and defined set of elements in an orderly way.
Coding is mostly done by means of symbols, (usually digits or letters), resulting in a concise representation.

Example: The assignment of code element: "CDG" as the concise representation of airport name: "Paris Charles de Gaulle". This airport name belongs to the set of airport names maintained by the International Air Transport Association (IATA). IATA has set the rules for establishing the concise representations.

Coding is a necessary tool for information processing. Coding of information enables it to be recorded, interpreted, processed and transmitted by humans and/or by machines.
All kinds of information can be coded: information about products, persons, processes, documents, countries, currencies, packages, etc.

Before making agreements on the coding methodology, i.e. the representation of information concerning events, actions, concrete or abstract objects in the real world, it must be investigated which data are relevant for the intended application. Information analysis of the universe of discourse concerned has to determine the role of the data in the information structure thereof.
In doing this a clear distinction should be made between identification, classification and reference needs.

5 Coding objectives

Information about any abstract or concrete object, action, or event of interest, (its characteristics or attributes) can be coded. Before making agreements on the configuration of their representation, the coding rules, it is necessary to determine the objective of the coding effort. It is not enough to design an ordered short representation of certain data. First the information requirements must be clear.

The following requirements generally occur:

- identification
- classification
- key to further information.

When data modelling is applied for the specification of application systems or messages for use in data interchange, the objectives of the users in the application environment will determine the choice of the entities and attributes to be taken into account as well as their interrelationship. The methods to be used for identification, classification or referencing will depend on those objectives.

5.1 Identification

The purpose of identification is to distinguish elements of a set from each other.
To be able to do this it must first be determined which characteristics have to be taken into account. Based on the selected characteristics comparisons can be performed and it can be ascertained whether an element of the set is equal to another element or different from it. To which degree of detail characteristics have to be recorded to indicate similarity or difference of elements of the set, depends on the area of application for which identification is needed.

<u>Example</u>
For its stock control of stationery an organization wants to identify various types of sheets of paper. Sometimes it may be sufficient to distinguish the various formats: A3, A4 or A5. Depending on the utilization of the sheets of paper other characteristics may be added, such as thickness, weight, chemical composition. If no specific requirements have to be met the recording of the format may suffice. In other cases, where handling and processing of the sheets of paper are critical the necessary characteristics have to be mentioned.

So, identification can be defined as:
The systematic registration of characteristics of elements of a set in such a way that they can be distinguished from each other.
The characteristics to be distinguished are inalienably part of an object or concept of interest.
The extent of the details to be observed is dependent on the user's objectives and the area of application.
These criteria result in the design or selection of an identification system.

<u>Examples of details to be distinguished:</u>

In the real world a person has a family name and a given name, is born in a country, on a certain date, resides in a country, his eyes have a certain colour, etc.

Application 1: In a governmental application system one wants to distinguish the colours of eyes of citizens.

Application 2: In an application system for medical research project one wants to distinguish the colours of eyes of human beings.

In these applications the required degree of discrimination need not be identical.
In the medical research application, it may be required that more colours of eyes of human beings are to be distinguished than in an application of a governmental body registering the colours of eyes of citizens.

In practice this may result into a compromise when selecting an identification system. Often the choice of a system is then determined on basis of the wish to have a minimum number of identification

systems to accommodate a maximum number of functions.

The objectives and applications determine which intrinsic characteristics will be taken into consideration.

5.2 Classification

The purpose of classification is to group objects or concepts of interest into classes in accordance with predetermined characteristics based on which similarities can be ascertained.
Classification is often used to support decision making or to get insight on trends or developments, without having to examine each instance of a set separately.

So, classification can be defined as:
A systematic arrangement of elements in groups or categories based on the similarity of predetermined characteristics.
Classification is done by means of control characteristics, i.e. those characteristics which have been assigned or are related to an object or concept of interest.
These characteristics may be intrinsic or extrinsic.

Example of control characteristics:
Place of manufacturing of products, turnover speed, market sector, production process for a product.

The information requirements and the business policy are determinative for the choice of control characteristics.

An organization may choose to apply various classification systems for the same type of objects, dependent on different needs.

Example:
A product may be classified according to
- function on behalf of sales
- manufacturing process on behalf of production
- value on behalf of inventory control
- volume/weight on behalf of transport
- type on behalf of Customs or statistical requirements

5.3 Key to further information

A key is an identifier within a set of data element types. Within the context of an application or data interchange a key shall be unique. In many application systems a reference number is needed as key to further information. The key in itself can be meaningless, but it gives access to the data required.

Examples:
- an order number may give access to the party to whom the order was sent, on which date, and the goods or services ordered;
- an article number may be related to a description, its price, its production process, the place of manufacturing;
- a salary number may refer to an employee, his name and address, his birthdate, his rank, his salary.

Reference numbers may be identifying in one application area, and classifying in another.

6 Types of codes

This chapter provides a description of basic coding methods. It is intended to assist in selecting appropriate code structures based upon specific application requirements and the nature of the elements in the set to be coded. It also provides principles and criteria to be considered in assessing alternative code structures, and mentions advantages and disadvantages of each coding method.

The choice of code structures is fairly extensive. The following information, however, should help to select the best method.

6.1 Forms of codes

The coding methods discussed in this chapter are outlined by the following listing. The set of methods shown is not exhaustive but does include all the significant types.
Many code structures applied in practice are often combinations of these basic types.

Non-significant codes
Sequential
  incremental sequential
  group sequential
  arranged sequential (chronological, alphabetical)
Random

Significant codes
Mnemonic
Abbreviation based
Matrix
Hierarchical
Juxtaposition
Combination
Value addition

6.2 Sequential codes

6.2.1 Principle

Elements of a set to be coded are assigned a number taken sequentially from an ordered set of numbers. These numbers are mostly natural integer numbers (e.g. beginning with "1") but alphabetic characters may also be used, e.g. AAA, AAB, AAC ...

Sequential assignment of code elements may be based on lists built in various ways, for example:

1. The list of natural integer numbers limited to the number of possibilities wished to be available.
2. Lists of numbers arranged on the basis of an algorithm, e.g. only even numbers or multiples of 10.