TECHNICAL REPORT

ISO/IEC TR 20943-1

First edition
2003-08-01

Information technology — Procedures for achieving metadata registry (MDR) content consistency —

Part 1: Data elements

Technologies de l'information — Procédures en vue d'obtenir la cohérence du contenu d'un registre de metadonnées (RM) —

Partie 1: Éléments de données

Reference number
ISO/IEC TR 20943-1:2003(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, the joint 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 of an agreement on an International Standard;

— type 3, when the joint 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 transformed into International Standards. 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.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 20943-1:2003, which is a Technical Report of type 3, was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 32, *Data management and interchange*.

ISO/IEC 20943 consists of the following parts, under the general title *Information technology — Procedures for achieving metadata registry (MDR) content consistency*:

Note: Parts 2 and 3 are currently under development.

— *Part 1: Data elements*

— *Part 2: XML structured data*

— *Part 3: Value domains*

Introduction

The exchange of metadata between metadata registries based on International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 11179 *Information technology — metadata registries (MDR)*, depends not only on registry software that conforms to the standard, but also on metadata contents that are comparable between registries. While the standard has provisions for data specification and registration, there are pragmatic issues pertaining to populating the registries with content. Based on the experiences of organizations that are implementing the standard, a Technical Report to explore content issues will help current and future users.

Metadata registries can be used to register data elements, value domains, and associated attributes for many kinds of organizational data resource collections. Metadata registries can store information on data elements used on forms, represented in enterprise data models, contained in EDI message sets, and described in documents and standards, as well as those data elements that are part of computer system applications. Some organizations use the registry to record essential facts about how data elements are used in existing applications, while other organizations use the registry as a repository of standard data elements to be used as models for data elements in application development. ISO/IEC 11179-6 specifically addresses the development and population of metadata registries.

ISO/IEC 11179-3 models a data element and its associated data element concept. Conceptualization and articulation of rules and relationships are needed in the creation of data element concepts, data elements, and value domains. Explication of the various possible levels of data elements and data element concepts and their relationships would greatly assist in the creation of shareable, well-formed data. Relationship and inheritance from the most generalized data element to the most specialized application data element need to be specified. Reuse of data value domains should be enabled and regularized.

While metadata registries can be used for storing information about a variety of metadata entities, this report addresses only data elements and associated metadata items. The goal of this paper is to ensure that there is a common understanding of the content of the data element attributes so that metadata can be shared between registries, despite their differences.

This Technical Report is based ISO/IEC 11179-3 of the six-part ISO/IEC standard that describes the organization of a registry for managing the semantics of data. The standard specifies the structure of a registry in the form of a conceptual model. The conceptual model is not intended to be a logical or physical data model for a computer system.

Information technology — Procedures for achieving metadata registry (MDR) content consistency —

Part 1: Data elements

1 Scope

1.1 Background

An ISO/IEC 11179-based metadata registry (MDR) (hereafter referred to as a "registry") is a tool for the management of shareable data; a comprehensive, authoritative source of reference information about data. It supports the standardization and harmonization processes by recording and disseminating data standards, which facilitates data sharing among organizations and users. It provides links to documents that refer to data elements and to information systems where data elements are used. When used in conjunction with an information database, the registry enables users to better understand the information obtained.

A registry does not contain data itself. It contains the metadata that is necessary to clearly describe, inventory, analyse, and classify data. It provides an understanding of the meaning, representation, and identification of units of data. The standard identifies the information elements that need to be available for determining the meaning of a data element (DE) to be shared between systems.

1.2 Purpose

The purpose of ISO/IEC TR 20943-1:2003 is to describe a set of procedures for the consistent registration of data elements and their attributes in a registry. ISO/IEC TR 20943-1:2003 is not a data entry manual, but a user's guide for conceptualizing a data element and its associated metadata items for the purpose of consistently establishing good quality data elements. An organization may adapt and/or add to these procedures as necessary.

1.3 Scope

The scope of ISO/IEC TR 20943-1:2003 is limited to the associated items of a data element: the data element identifier, names and definitions in particular contexts, and examples; data element concept; conceptual domain with its value meanings; and value domain with its permissible values.

1.4 Registration approach — data elements and value domains

There is a choice when registering code sets and other value domains in an ISO/IEC 11179 metadata registry. Some Registration Authorities treat these sets as value domains, and others treat them as data elements. For the purposes of ISO/IEC TR 20943-1:2003, the choice will always be to treat the sets as data elements unless explicitly stated. This choice is made to help illustrate the way to register many different kinds of data elements, including examples for registering standard code sets as data elements.

2 Normative references

The following referenced documents are indispensable for the application 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/IEC 11179-1:—[1)], *Information technology — Metadata registries (MDR) — Part 1: Framework for the specification and standardization of data elements*

ISO/IEC 11179-2:—[1)], *Information technology — Metadata registries (MDR) — Part 2: Classification for data elements*

ISO/IEC 11179-3:2003, *Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes*

ISO/IEC 11179-4:—[1)], *Information technology — Metadata registries (MDR) — Part 4: Rules and guidelines for the formulation of data definitions*

ISO/IEC 11179-5:—[1)], *Information technology — Metadata registries (MDR) — Part 5: Naming and identification principles for data elements*

ISO/IEC 11179-6:—[1)], *Information technology — Metadata registries (MDR) — Part 6: Registration of data elements*

ISO/IEC TR 15452:2000, *Information technology — Specification of data value domains*

Standards from which examples have been drawn to be used in this document are listed in the Bibliography.

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 11179 and ISO/IEC TR 15452 apply.

4 Data element abstraction

This clause presents a conceptual framework for structuring data elements and data element contents in a registry. Data elements are ideally the result of a process of development, involving several types of abstraction, producing a series of "layers" related to each other by the method of abstraction used to produce one from the other. Layers may progress from the more general to the more specific.

Depending on the type of abstraction, relationships among the members of each layer and between layers are meaningful in terms of defining the structure of the registry contents. This provides a means of comparison of the contents of different registries and of searching within a registry. In addition to the data element definition and other attributes, comparing the type and level of abstraction by which the data element was derived can ensure that content can be shared among registries.

One could use layers to structure development of a system, for instance, with the highest layers of definition contained in a business view, and development progressing to the implemented system layer. The number and granularity of layers are driven by user requirements. This clause will describe two ways to derive layers, neither of which are intended to be mandatory for any particular implementation, and will present examples of the types of abstraction most useful to registry implementations.

---

1) To be published.

4.1 Abstraction types

Abstraction is a well-developed tool for analysis and conceptualization. It is used as a way of focusing on parts of the model of interest to a particular process or function. The term "abstraction" is used to refer both to the process and the results of the process. Abstraction can be applied to the registry environment as a way to articulate the development of data elements and their relationships to each other.

Several methods can be used to achieve the decomposition of layers from the most abstract to the more concrete. Starting with the most general conceptual notions and progressing to the data elements in applications, these layers can be labeled by the type or types of abstraction used to produce them from another layer. Although this description of the process of abstraction development implies that it always proceeds in one direction (from the general to the concrete), there may certainly be cases where the reverse is true. Also, after the structures are in place, there is no implication that the relationships are all in single directions. Indeed, the names of the abstraction types deliberately include two levels in a effort to avoid that assumption.

The three types of abstraction of most interest to data element development are *specialization/generalization, concatenation/decomposition* and *aggregation.*

- **Specialization/generalization** is a relationship between two classes, where all items in one (subclass) are also in the other (superclass).

- **Concatenation (**or **composition)** involves the development of composite values by concatenation of character sequences from source values. **Decomposition** involves the separation out of the component portions of value sequences into their source character values.

- **Aggregation** involves the derivation of values by means of mathematical operations on source values. It is not usually possible to reverse the operation (called **derivation**) to recover the component parts.

All three types are commonly found in registries.

4.2 Example of specialization/generalization

The mechanism of specialization/generalization can be used to express a hierarchical structural relationship among data elements.

The United States Postal Service (USPS) establishes postal codes for the United States. The data element State USPS Code (figure 1) forms part of an address group. It might be used to map to application data elements directly, but it could also be the highest layer of a hierarchy of data elements, each displaying a greater degree of specialization according to its position in the structure.

The next layer down contains the elements Geographic Address State Code and Mailing Address State Code. The two State Code elements are now members of two differing address groups. The conceptual domains of the two elements are still identical, however, as the set of potentially valid values of geographic addresses and mailing addresses are the same.

Further specialization takes place at the next layer. Mailing Address State Code is partitioned into Facility Mailing Address State Code and Customer Mailing Address State Code. As before, the address groups of these elements differ, and in addition the data value domains do not necessarily contain the same sets of valid values. The domain of Customer Mailing Address State Code might remain the same as Mailing Address State Code, but the domain of Facility Mailing Address State Code might now be restricted to the subset of states in which facilities are located.

In this example, application data elements are mapped to the registry elements at this layer. Subclause 6.7 describes linking data elements that occur in different databases, under different names, to one registry data element. These relationships encourage data interchange by assuring semantic equivalency among disparately appearing data elements.

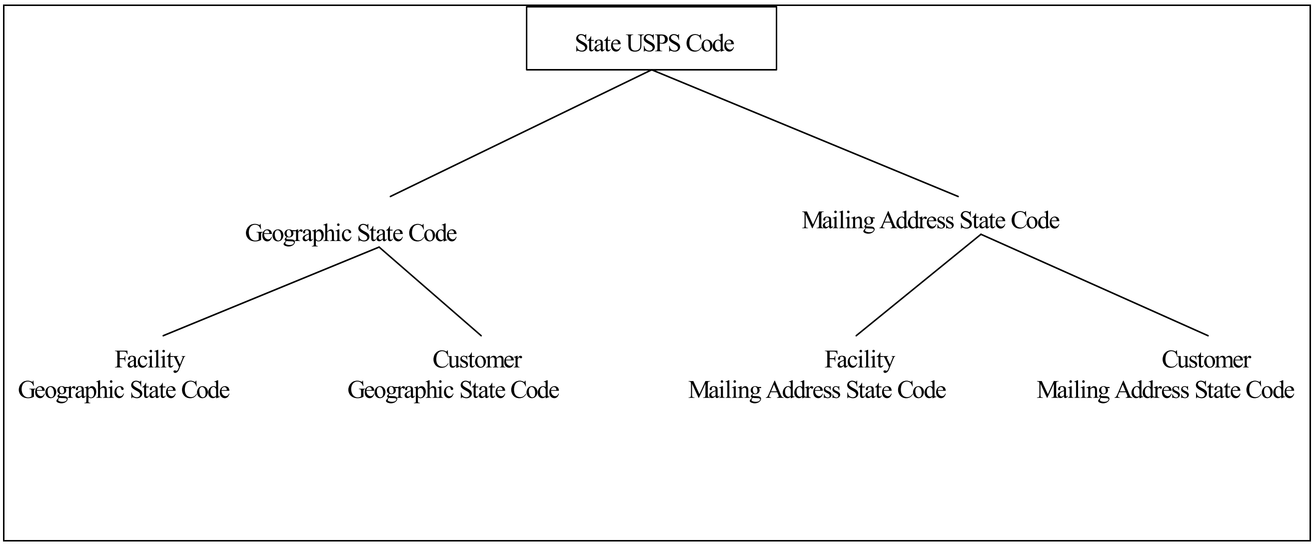

Figure 1 — Specialization/generalization hierarchy

Clauses 6, 7 and Annex A show a way to track the layers of specialization by assigning Layer of Abstraction Type as a value of Classification_Scheme_Item.csi_type_name and values such as specialization, generalization, aggregation, and concatenation to the attribute Classification_Scheme_Item.csi_value in the data element description. These attributes show the type of abstraction of the data element in the abstraction structure. In the example above, the attribute value for data element State USPS Code's Classification_Scheme_Item.csi_value is designated as Generalization. This designation indicates that this data element is at the top of the specialization/generalization structure for data elements in this registry. The data elements in the next layer, Geographic State Code and Mailing Address State Code, are assigned the attribute value Specialization. The data elements in the third layer, namely Facility Geographic State Code, Customer Geographic State Code, Facility Mailing Address State Code, and Customer Mailing Address State Code, are also assigned Specialization.

There may be other structures in the registry in which data elements are placed relative to other registry entries, such as data element concepts. In that case, other attributes may be needed to locate the example data elements in those structures.

4.2.1 Example of sharing a value domain

Figure 1 shows a specialization/generalization hierarchy where State USPS Code is viewed as a data element that is specialized as Geographic State Code and Mailing Address State Code. If the ranges of permissible values for Geographic State Code and Mailing Address State Code are identical, then an alternative representation of this scenario is to consider State USPS Code as a data value domain that is shared by the data elements Geographic State Code and Mailing Address State Code.

4.3 Example of concatenation/decomposition

Use concatenation/decomposition to express a relationship among data elements in which the higher layer describes a whole sequence, and the lower layers are component parts of that sequence. They may or may not be discrete; that is, some of the lower layer components may overlap in their descriptions or domains.

An example of this type of abstraction is the dividing of mailing address sequences into their subcomponents (figure 2). These divisions are based on ISO 11180, Postal Addressing. The figure shows some of the subdivisions for the data element group Mailing Address Group.

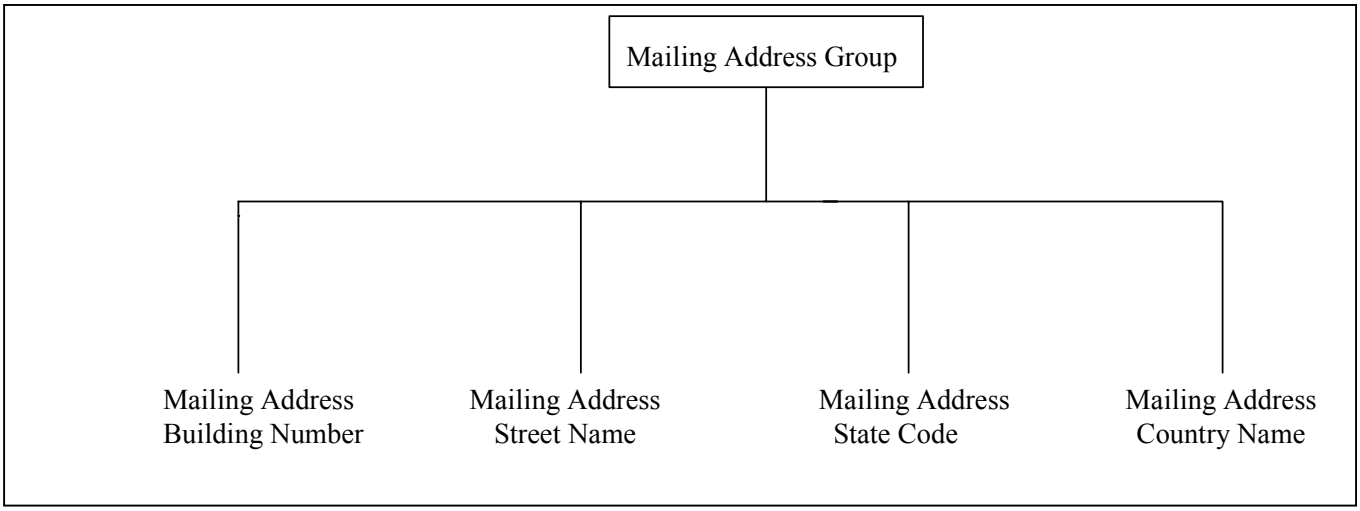

Figure 2 — Concatenation/decomposition example

4.4 Example of aggregation

Use aggregation to express a relationship among data elements in which the higher layer describes a characteristic of a whole and the lower layers are factors affecting that characteristic. They may or may not be discrete; that is, some of the lower layer components may overlap in their descriptions or domains.

Examples of this type of abstraction include:

- the derivation of a period of operation by subtraction of the commencement time, and any temporary stoppage periods, from the finishing time.
- the derivation of an average and standard deviation from a population of values and the count of those values.

It should be noted that it is not usually possible to reverse the aggregation operation (which is called derivation). This is because the derivation process produces a different type of information to the original data instances. In a strict analysis of derivation, information is usually lost in the process.

Clauses 6, 7 and Annex A show a way to track the layers of concatenation by assigning Layer of Abstraction Type as a value of Classification_Scheme_Item.csi_type_name and values such as specialization, generalization, aggregation, and concatenation to the attribute Classification_Scheme_Item.csi_value in the data element description. These attributes show the type of abstraction of the data element in the abstraction structure. In the example above, the attribute value for data element Mailing Address Group's Classification_Scheme_Item.csi_value is designated as Concatenation. This designation indicates that this data element group is at the top of the concatenation/decomposition structure for data elements in this registry. The data elements in the next layer of decomposition are assigned the attribute value Decomposition.

Throughout Clauses 6 and 7, reference will be made to the layer of abstraction for each example data element. The relationship of elements in any particular layer to those above and below it will help the user to choose which data element contains the appropriate information for a particular use.

There may be other structures in the registry in which data elements are placed relative to other registry entries, such as data element concepts. In that case, other attributes may be needed to locate the example data elements in those structures.

5 Data element registration

Registration of a data element in a registry requires that certain characteristics of the data element be recorded to clearly describe and define it. These characteristics are stored as attributes of the data element. A registry can be used to record information about a range of data elements, from those found in carefully crafted data standards to those found in applications. The amount and quality of metadata information available can vary from good, complete information to poor, incomplete information. This Technical Report is intended to describe the population of a registry with data elements for which good quality, consistent metadata can be created. ISO/IEC 11179-3 specifies attributes for recording information about a data element in a registry. This Technical Report gives examples that demonstrate the population of a registry. It includes attributes that are mandatory and fully defined by the metamodel, as well as those where the Registration Authority must establish its own profile of required attributes.

There are often problems associated with poorly-formed data elements. For instance, values with definitions other than those defined in the formal definition attribute may be mixed into a value domain along with those which are properly defined. Two data elements may contain overlapping values. This document does not address these kinds of problems.

The metamodel provides for the recording of administrative data about many metadata items of the registry. These are considered to be Administration records. Administration records must contain an identifier, the date the item was created, and the registration and administrative status of the item. Administration records must have a name and a definition in context. An administration record may also contain dates when data about the item were changed, an effective date, and an end date. Explanatory notes and the origin of the item, as well as administrative notes, change description, and unresolved issues may also be recorded. The exact process for determining registration status of administration records other than data element itself, has not been established in the current ISO/IEC 11179-6, where registration is described. For the purpose of this Technical Report, the characteristics of completeness and data quality that are to be used to determine registration status for all items is considered to be the same as for data elements themselves.

Two approaches to data element registration are included in this report: a bottom-up approach, where limited information about the data element is available prior to registration, and a top-down approach, where conceptual information has previously been identified.

- The bottom-up registration procedure provides for the basic metadata attributes about the data element (e.g., definition, name, and permissible values) to be completed prior to defining the conceptual information about the data element. This approach is described in Clause 6 of the Technical Report.

- A top-down approach is useful in many circumstances. Although it requires more "up front" effort, top-down registration has the potential to produce more stable and uniform metadata. An example of a top-down registration, where registration begins with identification of data element concepts, is provided in Clause 7.

It should be noted that the same rules and guidelines apply to the associated metadata items and attributes of a data element, regardless of the approach. The bottom-up and top-down procedures differ only in the order that the practitioner analyzes the data element and formulates its associated items and attributes. The complete registration process is described in context in Clause 6. Except for the order described, the same processes are applicable in Clause 7.

6 Bottom-up approach to data element registration

In many cases where a data element is submitted for registration, only a limited amount of information (e.g., a name, definition, and a set of permissible values) is provided by the submitter. Other attributes must be determined based on an understanding of the underlying data values and concepts that are implied by those facts. Registry practitioners, working from existing systems, may find that working from data elements to data element concepts is appropriate. This approach is referred to as bottom-up in this Technical Report. A bottom-up approach might also be used where the registry is intended to serve as a distribution mechanism for metadata that describes the data in data products such as public data sets, query results, etc.

The examples provided in this clause describe how to formulate attributes about a data element, based on a bottom-up procedure. The registration procedures are presented in a logical order for analysing and formulating attributes for a data element. First a general procedure for registering data elements is described, followed by examples of registration of four types of data elements, including data elements from:

- an International Standard with an enumerated domain;
- an information system, where the application data element uses that standard enumerated domain;
- an International Standard with a non-enumerated domain;
- another standard, where the application data element uses that standard non-enumerated domain.

This Technical Report is intended to be used to help registry practitioners to formulate the attributes that describe and define a data element. Subclause 6.1 presents an example of an overall approach to bottom-up data element registration. Subclauses 6.2, 6.3, 6.4, and 6.5 should be consulted for more specific examples of registering the kinds of data elements described in International Standards, national standards, and in information systems. Annex A contains a table that concisely summarizes the information registered for each data element in the examples that follow in Clause 6. Annex B, which is based on ISO/IEC 11179-2, -4, -5 and -6, contains more detailed information and examples to assist the practitioner who is registering data elements. Annex C provides a crosswalk from the example attributes named in this Technical Report to the ISO/IEC 11179-3 metamodel.

6.1 General procedures for registering a data element

Often only a limited amount of information is available about a data element that has been submitted for registration, e.g., the name and definition contained in a document or provided by the submitting organization and a set of permissible values, where appropriate. The general procedures that follow are intended to result in the registration of a complete, well-defined data element that meets the requirements of a particular registration authority.

It should be noted that the metadata for some data elements in a registry will never be complete. This is true of application data elements that are obtained from computer software, where very little information is known except the representational attributes (e.g., field length and datatype). For these data elements, only the most basic attributes will be entered, and the data element's registration status will remain incomplete.

6.1.1 Understanding the data element

When examining existing computer systems, the practitioner should not automatically assume that there will be a one-to-one correspondence between a field in a record and a data element in the registry. Often as systems evolve, fields become used for multiple purposes under various conditions. When such a situation is detected, it may be desirable to establish more than one data element, corresponding to a single field. The reverse situation where multiple fields correspond to a single data element is also possible, though less likely.

The first step in the registration procedure is to gain an understanding of the data element. What kind of data will be stored in this data element? Is there a definition or description of the data values? Were permissible values or examples of the data provided? Will the data values be determined by an arithmetic or statistical procedure? What will the data values look like; e.g., are they names or descriptions of things, numerals to be calculated, strings of characters and numbers that are identifiers? Where documentation is inadequate to fully understand the data element, the practitioner must consult those who represent the source of the data element to obtain the necessary information.

The result of this first step is an understanding of the semantic content of the data element.

6.1.2 Content research

Prior to formulating attributes toward registration of a new data element, the registrar should perform content research to determine the following: