INTERNATIONAL STANDARD

ISO 12620

Third edition
2019-05

Management of terminology resources — Data category specifications

Gestion des ressources terminologiques — Spécifications des catégories de données

Reference number
ISO 12620:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
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 on 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 the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 3, *Management of terminology resources*.

This third edition cancels and replaces the second edition (ISO 12620:2009), which has been technically revised.

The main changes compared to the previous edition are as follows.

ISO 12620:2009, *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*, described a data model and management features for a Data Category Registry designed for the purpose of standardizing data category specifications. The current edition of ISO 12620 has been streamlined to eliminate the standardization function previously built into the data model. It describes requirements for maintaining a consensus-based, industry-appropriate repository of harmonized data category specifications for use in language resources.

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

Data associated with language resources are identified, collected, managed and stored in a wide variety of environments. Data appearing in language resources are generalized into classes that are referred to as *data categories*. Differences in approach for developing different kinds of language resources as well as differences in technical environments inevitably lead to variations in data category definitions and data category names. The use of uniform data category names and definitions employed in resources within the same linguistic domain (for example, among terminological resources, lexicographical resources, annotated text corpora, etc.) contributes to system coherence and enhances the re-usability of data. Such uniform use requires access to formal data category specifications. Defining a clear framework for specifying, managing and using data categories will increase interoperability of language resources.

Management of terminology resources — Data category specifications

1 Scope

This document provides guidelines and requirements governing data category specifications for language resources. It specifies mechanisms for creating, documenting, harmonizing and maintaining data category specifications in a data category repository. It also describes the structure and content of data category specifications. The intended audience of this document is researchers and practitioners in fields of language resource management who use data categories and data category specifications.

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 24619, *Language resource management — Persistent identification and sustainable access (PISA)*

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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/

3.1
conceptual domain
permissible content of a *data category* (3.2)

EXAMPLE In a terminology database, the data category /part of speech/ could have a conceptual domain consisting of the values: /noun/, /verb/, /adjective/, /adverb/.

Note 1 to entry: The permissible content can be enumerated (such as in a pick list), as in the example, or subject to formal restrictions such as dates, or free text such as the conceptual domain of /definition/. Although the latter type is not formally restricted, it is nevertheless subject to adherence to the requirements of its data category specification, i.e., it contains a true definition and not a note, example, or some other piece of information.

3.1.1
open conceptual domain
conceptual domain (3.1) that has no formal restrictions

Note 1 to entry: An open conceptual domain is frequently associated with data categories that take free text as their content, such as /definition/.

Note 2 to entry: Some requirements are not machine processable, for instance, to require that /definition/ only contain definitional information.

3.1.2
closed conceptual domain
conceptual domain (3.1) that is restricted to a set of enumerated values

EXAMPLE The data category /grammatical gender/ can have a conceptual domain consisting of the values

/feminine/, /masculine/ and /neuter/.

3.1.3
constrained conceptual domain
conceptual domain (3.1) that is restricted to a constraint or rule specified in a schema-specific language

EXAMPLE The data category /date/ can be constrained by a system setting to certain date formats, or a data category can be subject to a termbase-specific rule, such as making it mandatory to enter a /source/ for a /definition/.

3.1.4
simple conceptual domain
conceptual domain (3.1) that has only two values

Note 1 to entry: The two values can be "yes" or "no", "true" or "false", or other such binary representation.

3.2
data category
DC
class of data items that are closely related from a formal or semantic point of view

EXAMPLE /part of speech/, /subject field/, /definition/.

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

Note 2 to entry: In running text, such as in this document, data category names are enclosed in forward slashes (e.g. /part of speech/).

[SOURCE: ISO 30042:2019, 3.8]

3.2.1
open data category
data category (3.2) that has an *open conceptual domain* (3.1.1)

3.2.2
closed data category
data category (3.2) that has a *closed conceptual domain* (3.1.2)

3.2.3
constrained data category
data category (3.2) that has a *constrained conceptual domain* (3.1.3)

3.2.4
simple data category
data category (3.2) that has a *simple conceptual domain* (3.1.4)

Note 1 to entry: See also *pick list value* (3.9).

3.3
data category concept
semantic content of a *data category* (3.2), independent of any specific implementations

3.4
data category name
linguistic representation of a *data category* (3.2) as it appears in a particular language or in a particular application or language resource

EXAMPLE The data category name for /part of speech/ is "part of speech" in English, and "partie du discours" in French.

3.5
data category repository
DCR
digital repository of *data category specifications* (3.7)

Note 1 to entry: Data category repositories are used as references when specifying language resources.

Note 2 to entry: A DCR for language resources is available at www.datcatinfo.net.

3.6
data category selection
DC selection
set of *data category specifications* (3.7) selected from a *data category repository* (3.5)

Note 1 to entry: A data category selection can represent the *data categories* (3.2) used within a research discipline or a specific application or project.

3.7
data category specification
DC specification
complete descriptive record of a *data category* (3.2)

3.8
persistent identifier
PID
unique Uniform Resource Identifier (URI) that provides permanent access to a digital object independently of its physical location or current ownership

EXAMPLE http://www.datcatinfo.net/datcat/DC-70.

[SOURCE: ISO 24619:2011, 3.2.4, modified — order of terms inverted, definition slightly reworded, note deleted, example added.]

3.9
pick list value
one of the enumerated or permissible values of a *closed data category* (3.2.2)

EXAMPLE "singular" and "plural" are pick list values in a field labelled "Grammatical Number".

Note 1 to entry: See also *simple data category* (3.2.4).

Note 2 to entry: Due to data modelling variance, most types of information that can be represented as pick list values in a database can also be represented as simple data categories. For example, "Plural" can be implemented as a checkbox, which, when checked, takes the value "yes" and when unchecked, takes the value "no".

4 Data categories and data category specifications

A data category (DC) is a class of information that forms part of a data collection or annotation scheme for a given language resource. For example, /definition/ and /part of speech/ are common data categories in terminological and lexicographical resources. Data category names can appear as the name of a field in the user interface of a software application, or as a markup element in an annotated resource.

Some data categories are pertinent to a specific application, research area, or type of resource and not others. For instance, a /concept identifier/ is characteristic of terminological or ontological resources, whereas /sense number/ is applicable to lexicographical resources. On the other hand, many data categories, frequently those of a strictly linguistic nature such as /part of speech/, /grammatical gender/ and /grammatical number/, are common to a wide variety of resources. These data categories may not always be implemented in the same way in different resources or applications, but each nevertheless evokes one universal data category concept. For instance, for terminology management, only a small

set of values are needed for /part of speech/ (e.g. noun, verb, adjective, adverb), but in lexicographical resources, many additional values are required (e.g. preposition, pronoun, etc.).

A data category specification (DC specification) provides the complete and formal representation of a data category (for example, its name, definition, examples, comments, etc.). Data category specifications can be referenced by the language resources that use them, for instance, through the use of persistent identifiers that directly resolve to the data category specification from within that resource.

5 Data category repositories

Data category specifications are normally stored in electronic format in a specially-designed database. This database is called a data category repository (DCR). Today, it is essential for DCRs to be available on the internet. For instance, a DCR for language resource descriptions, named DatCatInfo, is available at www.datcatinfo.net (see Annex A).

Researchers and software developers working with language resources benefit greatly from being able to access a trusted source of information about data categories. Providing a precise description of the data categories that are used within a given data collection allows for a quick diagnosis of its compatibility with other data collections or its suitability for use in computer processes. A DCR containing vetted data category specifications provides users with the information they need to implement data categories in a manner that is consistent with other users. Consequently, the interoperability of language resources is greatly enhanced.

Interoperability of language resources is a key factor for supporting innovation and progress in various focus areas of the language industry, such as terminology management, natural language processing, and annotation schemes. These areas support important sectors of our economy and social development such as global communication and trade, knowledge extraction, and content management.

To support research and development in language resources, it should be possible for users of a DCR to subset collections of data category specifications from the whole DCR for application-specific uses. These subsets are referred to as data category selections (DC selections). A data category selection defines, in combination with a data model and optionally additional constraints, a given application-specific language resource. For example, according to ISO 30042, a selection of terminology-related data category specifications, together with the metamodel defined in ISO 16642 and some additional specified constraints, constitutes a TBX data category module. Modules are combined to define a TBX dialect, which is a TML (Terminology Markup Language) as described in ISO 16642.

6 Requirements for data category specifications

This clause states the requirements that data category specifications shall fulfil in order to support the effective use of data categories for language resources.

A data category specification shall:

— be available online;

— provide a unique mnemonic identifier of the data category;

— document the various acceptable names of the data category, in different languages and for various applications where desired;

— provide a clear definition of the data category concept, in different languages where desired;

— indicate the content model of the data category: the types of information that the data category allows when implemented. For instance, the data category /grammatical gender/ might only allow a limited set of values such as /masculine/ and /feminine/, whereas the data category /definition/ allows free text;

— describe how the data category is implemented and used in:

  — specific projects or initiatives;

  — specific types of language resources;

  — specific languages or linguistic or cultural contexts;

  — specific sub-domains of language resources where the data category is relevant;

— describe how the data category is represented in various annotation schemes and markup languages;

— include administrative information, i.e. dates and user names, to track the creation and modification of the data category specification;

— include information indicating its stage in a vetting process, such as: proposed, under review, approved, deprecated;

— include a historical record of changes to the data category specification;

— have a unique persistent identifier allowing it to be accessed directly from within an application or a language resource.

7 Requirements for documenting data categories

7.1 Identifiers and names

7.1.1 Data category specifications

Data categories specified in a data category repository shall be assigned the set of identifiers and names specified in 7.1.2 to 7.1.5.

7.1.2 A unique and stable mnemonic identifier

Each data category shall have a unique mnemonic identifier, which shall not include space characters for multi-word forms. As a consequence, camel case style, which involves capitalizing the first letter of each word after the first word in the identifier (see the example below), is recommended to maximize both human and machine-readability. These identifiers are used in encoding environments as elements or as attribute values.

EXAMPLE partOfSpeech.

7.1.3 A unique and persistent identifier (PID)

Each data category shall also have a unique and persistent URI identifier (PID), as per ISO 24619, which provides direct web access to its complete DC specification. PIDs provide a way of locating a resource and ensure that unique names and identifiers are associated with resources in the context of internet-based namespaces.

EXAMPLE www.datcatinfo.net/datcat/DC-396 (this is the PID for /part of speech/ in the DatCatInfo DCR).

7.1.4 A unique canonical data category name

Aside from unique mnemonic and persistent identifiers, which are meant to be machine-readable, data categories also need to have human-readable names for use in discourse. Each data category shall be assigned a name in a language that is selected as the main human-readable language of the DCR. This name, known as the canonical data category name, can be written according to standard spelling and

punctuation. Canonical data category names should be unique across the entire DCR, although during periods of harmonization this may not always be possible.

EXAMPLE part of speech, from the DatCatInfo DCR, where all canonical data category names are in English.

7.1.5 Language-specific data category names

In addition to the canonical data category name, names in other languages are permitted, and they can also be written according to standard spelling and punctuation of those languages.

The language-specific names are frequently used as field names or values in language resources and can therefore vary from application to application depending on computing environments or other constraints. For purposes of exchange or interoperability, variant data category names in a language resource shall be mapped to stable identifiers in the DCR, such as mnemonic or persistent identifiers.

EXAMPLE pos, word class, grammatical category (en)

partie du discours, catégorie grammaticale, classe du mot (fr)

Wortklasse, Wortart (de)

7.2 Conceptual domains and data category types

When data categories are implemented in software applications or language resources, they often have certain constraints on the types of information they can contain or how they otherwise behave within that environment. These constraints are often referred to as the "conceptual domain" of the data category. For instance, /date/ can only allow certain date formats, and /subject field/ might only allow certain predefined values. It shall be possible to clearly indicate the conceptual domain of each data category.

It is possible for a data category to require more than one conceptual domain in order to address different needs. For instance, the number of permissible values of /part of speech/ in morphosyntax research is much greater than that for terminology management. The DCR shall adopt one of the following two methods for handling this situation:

— allow more than one conceptual domain in a single data category specification; or

— require two separate data categories, one for each conceptual domain.

Each data category shall have one of the following conceptual domains:

— open conceptual domain;

— closed conceptual domain;

— constrained conceptual domain;

— simple conceptual domain.

The conceptual domain of a data category is an essential property that can be used to distinguish between different types of data categories to facilitate their use for data modelling. Simply speaking, an open data category is one with an open conceptual domain, a closed data category is one with a closed conceptual domain, a constrained data category is one with a constrained conceptual domain, and a simple data category is one with a simple conceptual domain.

The permissible values of a closed data category are simple data categories; they are often referred to as pick list values. To maximize interoperability of pick list values, they shall also have their own specification in the DCR.

7.3 Data elementarity

Data category specifications shall adhere to the principle of data elementarity, whereby a field within the specification shall only be used for its intended purpose. For example, it is important to clearly distinguish between the various descriptive fields such as definitions, explanations, examples, usage notes and comments. Putting an explanation in a definition field, or putting the source of a definition in the definition field, is an example of how this principle sometimes is violated.

7.4 Subsetting

As with any type of record that is added to a database, data category specifications shall be assigned to one or more logical categories so that they can be easily searched, retrieved and utilized for various purposes. For example, the datcatinfo DCR comprises categories based on communities of practice within the broad field of linguistics, such as terminology and lexicology. These categories shall be defined according to the needs of the end users (which can include software and data processing applications in addition to people). Categories that are based on semantic divisions of the overall domain of the DCR are also recommended as they help to clarify the meaning of data categories. Both types (user-needs based and semantics based) can be offered in the same DCR.

8 Requirements for a data category repository

This clause specifies the requirements that a data category repository (DCR) shall fulfil in order to support the effective use of data categories for language resources.

A data category repository shall:

— be available online;

— provide a collection of data category specifications for reference;

— provide an automated mechanism to avoid the creation of multiple specifications containing the same data category name;

— provide a mechanism whereby users can submit new data category specifications and provide feedback on existing data category specifications;

— provide search filters allowing subsets of the DCR to be searched and retrieved using various search criteria (for instance by date, by creator, by the content of a field, etc.);

— provide user access controls to limit write access to authorized DCR managers;

— allow data category selections to be defined for various applications or user groups;

— subset the data category specifications based on a rigorous ontology of the data category concepts;

— allow export of data category selections, e.g. to a CSV file or XML file.

9 Referencing data categories

The explicit reference to a data category shall be made by embedding the persistent identifier (PID) for its data category specification in the referencing resource. The PID is automatically assigned by the DCR. For instance, /part of speech/ can be referenced by a URI such as

www.datcatinfo.net/datcat/DC-396

Some schema languages have built-in constructs for embedding these PIDs. For instance, the following markup will signal that the element being specified (<pos>) has the meaning defined for /part of speech/ in the DatCatInfo DCR:

```
<elementSpec ident="pos">
<equiv name="partOfSpeech" uri="http://www.datcatinfo.net/datcat/DC-396"/>
</elementSpec>
```

Schema languages that lack these provisions, but which are still based on an XML vocabulary, can still embed the PIDs. For instance, in a Relax NG Schema, one could specify that a POS element is equivalent to/part of speech/ in the DatCatInfo DCR by embedding the dcr:datcat attribute at the appropriate location:

```
<rng:element name="POS" dcr:datcat="http://www.datcatinfo.net/datcat/DC-396">
</rng:element>
```

10 Harmonizing and vetting data categories

Although a mechanism is provided in the DCR to prevent the creation of data category specifications with identical data category names, when multiple people are involved in creating data category specifications, duplicate data category specifications can still occur in a DCR. A duplicate data category specification is one that refers to the same data category concept as another data category specification. Duplicate data categories and their specifications shall be identified and removed from the DCR. The process involved is referred to as harmonization.

Duplicate data category specifications do not necessarily contain the same data category names. There can be minor differences, which are relatively easy to detect, such as /grammatical category/ and /grammar category/. But there can also be duplicates that show no similarities at all. For example, /part of speech/, /grammatical category/ and /word class/ might all refer to the same data category concept.

Harmonization shall be carried out with a focus on three types of potential duplicates, in this order:

— data category specifications that have identical data category names;

— data category specifications that have similar data category names;

— data category specifications that have dissimilar data category names.

Harmonization shall be carried out in the following steps:

— identification of potentially duplicate data category specifications;

— comparison of the information that describes the data category concept for each set of potential duplicates. Resolution shall be carried out as follows:

  — if the data category concept is different, the data category specifications are not duplicates, and they shall be assigned unique IDs and PIDs;

  — if the data category concept and the conceptual domain are the same, the data category specifications shall be marked as duplicate, linked, or merged;

— a reviewer comment shall be added to the data category specification in order to provide a record of decisions.

No information describing the use of a data category for a specific application or user group shall be deleted or compromised during the harmonization process.

Duplicate data category specifications shall not be physically deleted from the DCR during the harmonization process, as this would remove any record of the harmonization. Instead, they shall

be separated from the remaining data category specifications. This separation can be achieved by assigning an identifying marker to the data category specifications in question so that they can be filtered to hide them from users during standard operations. Another method is to move them to a designated section of the DCR. These aforementioned methods eliminate the need to physically delete duplicate records from the DCR while the harmonization process is in progress. When harmonization is complete or well-advanced, and the DCR is considered to be in a stable state, it can be decided that some records can be deleted. In this case, those records shall be exported from the DCR to a file beforehand and archived for future reference.

Harmonization shall be carried out on a regular basis; the necessary frequency depends on how many people are involved and how often new data category specifications are created.

Vetting refers to the process of reviewing a data category specification and assigning it a status value that reflects its level of reliability and acceptability. As described in Clause 6, status values shall be available in the data category specification data model (proposed, under review, approved, deprecated). Vetting shall be carried out in consultation with the relevant stakeholder group.

11 Management

Data category specifications, and the DCR in which they are stored, shall be subject to clear and well documented management procedures. Assuming that the DCR is on a web site, documentation of these procedures shall be available on the same web site. The stakeholder group establishing a DCR shall appoint a board of experts to oversee overall governance. Technical support shall be provided for the DCR management software and for the hosting web server.

Specific individuals shall be appointed to harmonize data category specifications. These persons should have a suitable background and experience in areas covered by the DCR and should be provided with appropriate training on the specification of data categories.

Annex A
(informative)

The DatCatInfo Data Category Repository

A.1 General information

DatCatInfo is a DCR for language resources. It is available at the web site www.datcatinfo.net. This DCR contains data category specifications for data categories that are used in standards, such as ISO 30042 (TBX) or ISO 24613 (LMF), or in related industry reports and best practices. It is designed to meet all the requirements specified in this document. Information about data categories can be recorded in different languages. The DCR also provides a canonical name section for providing a data category name in the main working language of the DCR.

A.2 Data model

The data model of the DCR is similar to the metamodel for terminological resources specified in ISO 16642. Each entry has three structural levels (see Table A.1):

— **Concept level**

  Information provided at this level describes the data category concept independent of any specific language or implementation.

— **Language level**

  A language level is provided for each language supported in the DCR.

— **Data category name level**

  This level contains the data category name plus a set of fields describing the data category as it is associated with that name.

Administrative information (creator, creation date, modifier, modification date) is recorded at the concept level and the data category name level.