INTERNATIONAL STANDARD

ISO 12620-1

First edition
2022-07

Management of terminology resources — Data categories —

Part 1: Specifications

Gestion des ressources terminologiques — Catégories de données —

Partie 1: Spécifications

Reference number
ISO 12620-1:2022(E)

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

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

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

This first edition of ISO 12620-1, together with ISO 12620-2:2022, cancels and replaces ISO 12620:2019, which has been divided into parts and technically revised. The main changes are as follows:

- ISO 12620:2019 described procedures for defining data categories used in language resources and described requirements for maintaining a pragmatic, consensus-based repository of harmonized data category specifications for use in language resources. This document has been narrowed to focus on the structure and rationale associated with data category specifications per se.
- The sections of ISO 12620:2019 that dealt with the creation and maintenance of data category repositories have been moved to ISO 12620-2.

A list of all parts in the ISO 12620 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

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 (e.g. among terminology resources, lexical resources, annotated text corpora) 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.

The intended audience of this document is researchers and practitioners in fields of language resource management who use data categories and data category specifications.

Management of terminology resources — Data categories —

Part 1: Specifications

1 Scope

This document provides requirements and recommendations governing data category specifications for language resources. It specifies mechanisms for creating, documenting, harmonizing and maintaining data category specifications in a data category repository (DCR). It also describes the structure and content of 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 12620-2, *Management of terminology resources — Data categories — Part 2: Repositories*

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 terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://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/ can have a conceptual domain consisting of the values /noun/, /verb/, /adjective/, /adverb/.

Note 1 to entry: The permissible content can be closed, 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/ or /context/.

Note 2 to entry: Some requirements are not always machine-processable, for instance, to require that /definition/ only contain definitional information, or that a /context/ meet certain specified requirements.

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

EXAMPLE In a specific terminology database, the data category /grammatical gender/ can, for instance, have 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 binary values

Note 1 to entry: Each declared *picklist value* (3.10) can be implemented as a *simple data category* (3.2.4) with a simple conceptual domain.

Note 2 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* (3.4) are enclosed in forward slashes (e.g. /part of speech/).

[SOURCE: ISO 30042:2019, 3.8, modified — The admitted term "DC" has been added.]

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 *picklist value* (3.10).

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, in a particular application or in a 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 specification
DC specification
complete descriptive record of a *data category* (3.2)

3.6
data category repository
DCR
digital collection of *data category specifications* (3.5)

EXAMPLE DatCatInfo, a DCR for language resources (see Reference [4]).

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

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

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.

3.8
harmonization
<data categories> analysis and resolution of minor discrepancies between or among multiple *data category specifications* (3.5) treating the same *data category concept* (3.3)

Note 1 to entry: The aim of harmonization can be to merge duplicate or quasi-duplicate specifications into a single entry.

3.9
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 `https://datcatinfo.termweb.eu/datcat/DC70`

[SOURCE: ISO 24619:2011, 3.2.4, modified — The order of terms has been inverted, "uniform resource identifier (URI) that provides permanent access to a digital object" has replaced "identifier that ensures permanent access for a digital object by providing access to it" in the definition, the note to entry has been deleted and the example has been added.]

3.10
picklist value
one of the enumerated or permissible values of a *closed data category* (3.2.2)

EXAMPLE /singular/ and /plural/ as picklist values of the closed data category /grammatical number/.

Note 1 to entry: Due to data modelling variance, most types of information that can be represented as picklist values in a database can also be represented as *simple data categories* (3.2.4). For instance, /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 instance, /definition/ and /part of speech/ are common data categories in terminology resources and lexical 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 discipline or type of resource and not others. For instance, /concept identifier/ is characteristic of terminology resources or ontological resources, whereas /sense number/ is applicable to lexical 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 are not always 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 lexical resources, additional values are required (e.g. preposition, pronoun).

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

5 General recommendations for data category specifications

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

A data category specification should:

- 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, i.e. the types of information that the data category allows when implemented;

  EXAMPLE The data category /grammatical gender/ can be configured to 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, e.g. proposed, under review, approved, deprecated;

- include a historical record of changes to the data category specification;
- have a unique PID allowing it to be accessed directly from within an application or a language resource.

6 Detailed requirements for documenting a data category in a DCR

6.1 Identifiers and names

6.1.1 A unique and stable mnemonic identifier

Each data category in a data category repository (DCR) 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 as in 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

6.1.2 A persistent identifier (PID)

Each data category in a DCR shall also have a persistent identifier (PID), which is a unique URI in accordance with ISO 24619 and which provides direct web access to its complete data category 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 `datcatinfo.termweb.eu/datcat/DC396` (PID for /part of speech/ in DatCatInfo)

6.1.3 A unique canonical data category name

In addition to the unique mnemonic PIDs, which are meant to be machine-readable, each data category in a DCR shall have a human-readable name 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 this is not always possible during periods of harmonization.

EXAMPLE “Part of speech” is the canonical data category name for /part of speech/ from DatCatInfo, where all canonical data category names are in English.

6.1.4 Language-specific data category names

In addition to the canonical data category name, names in other languages are permitted. 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 identifiers or PIDs.

EXAMPLE

- pos, word class, grammatical category (en)
- catégorie grammaticale, partie du discours, classe du mot (fr)
- Wortklasse, Wortart (de)

6.2 Conceptual domains, data category selections and data category types

When data categories are implemented in software applications or language resources, the data categories for the application or resource are collected from the DCR to form the data category selection associated with that application. Data categories in a given data category selection often have certain constraints on the types of information they can contain or how they otherwise behave within that respective environment. These constraints are referred to as the "conceptual domain" of the data category. For instance, /date/ can only allow certain date formats, and /subject field/ can only allow certain pre-defined values. It shall be possible to clearly indicate the conceptual domain of each data category.

In some environments, a given data category can require more than one conceptual domain in order to address multiple needs. For instance, as noted in Clause 4, 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;
- require two separate data categories, one for each conceptual domain.

Each data category shall have one of the following types of 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. 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 often comprise "picklist values" and can be implemented as simple data categories. These picklist values should be treated in their own specifications in the DCR.

6.3 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 instance, 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 is sometimes violated.

6.4 Profiles

Data category specifications should be assigned to one or more logical groups so that they can be easily searched, retrieved and utilized as subsets. In this model, these subsets are called "profiles", which are based on (sub-)communities of practice within the broad field of language resources, such as terminology and lexicology. These data category profiles should be defined according to the needs of the end users, which can include automated software and data processing applications in addition to human examination. Indicating specific profiles is also recommended to clarify the significance of individual data categories. Data categories with the same name can be differentiated based on user-needs or semantic criteria. The profiles or classification systems specified for purposes of filtering or retrieving data category specifications in a DCR can become the principle on which data category profiles are defined, as documented in ISO 12620-2:2022, 4.2.

7 Referencing data categories

The explicit reference to a data category shall be made by embedding the 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:

```
datcatinfo.termweb.eu/datcat/DC396
```

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

```
<fs>
 <f name="POS"
  dcr:datcat="https://datcatinfo.termweb.eu/datcat/DC-396"/>
</fs>
```

Markup 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, it is possible to 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="https://datcatinfo.termweb.eu/datcat/DC-396"/>
</rng:element>
```

8 Harmonizing and vetting data categories

A mechanism should be provided in a DCR to prevent the creation of data category specifications with identical data category names. Nevertheless, 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, harmonized, and remaining duplicates removed from the DCR.

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 instance, /part of speech/, /grammatical category/ and /word class/ can 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 of duplications 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 duplicates, edited if necessary, or 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 involves the process of reviewing a data category specification and assigning it a status value that reflects its level of reliability and acceptability. Status values shall be available in the data category specification data model (e.g. proposed, under review, approved, withdrawn). Vetting shall be carried out in consultation with the relevant user group.

9 Management

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 customary for DCRs to be available on the internet or on local intranets. For instance, a DCR for language resource descriptions, named “DatCatInfo”, is available at Reference [4] (see Annex A). ISO 12620-2 specifies procedures for creating, managing and maintaining data category repositories. These procedures shall be followed.

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 website, documentation of these procedures shall be available on the same website. The community of practice 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 individuals 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)

Structure of a data category specification

A.1 General information

DatCatInfo is a DCR for language resources. It is available at Reference [4]. This DCR contains data category specifications for data categories that are used in ISO/TC 37 documents, such as ISO 30042 (TBX)[3] or the ISO 24613 series (LMF)[2], 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 data category specification 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 for specifying data categories is similar to the meta-model for terminology resources specified in ISO 16642[1]. Each entry has three structural levels (see Table A.1):

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

b) Language level: A language level is provided for each language supported in the DCR.

c) 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.

Table A.1 — Data model of the DatCatInfo DCR

Descriptor (field name)	Content	Description
Concept level		
Relations	Pointer to another data category	Includes hyperlinks that relate closed data categories to the picklist values that make up the conceptual domain. Relations are bidirectional, so that there is also a pointer from the picklist values to their corresponding closed data category.
Implemented as	Single-value selector: — picklist — text (free text) — picklist value	Indicates how this information is normally implemented in a language resource; “picklist” refers to closed data categories, “text” refers to open data categories, and “picklist value” refers to simple data categories. There are no constrained data categories in this DCR. Only one value is allowed. For instance: /part of speech/ – type: picklist /definition/ – type: text /noun/ – type: picklist value
PID	Free text	Non-mnemonic persistent identifier
Identifier	Free text	Unique mnemonic identifier

Table A.1 *(continued)*

Descriptor (field name)	Content	Description
Definition	Free text	A definition of the data category concept
Justification	Free text	Reason that justifies the creation of this data category specification
Origin	Free text	Source of this data category
Explanatory comment	Free text	Any type of comment that further explains this data category
Profile	Multi-value selector: — Metadata — Morphosyntax — Semantic content representation — Syntax — Language codes — Terminology — Lexicography — Sign language — Translation — …	Indicates the sub-field within the broader field of language resources where this data category is used or to which it applies. Multiple values are possible.
Language level		
Language	Single-value selector	A variety of languages are available. Permissible values are identified in IETF BCP 47[5].
Data category name level		
Data category name	Free text	One or more names of the data category. Multiple instances are possible, and each instance is given all the fields in the data category name level. At least one name in English is mandatory in DatCatInfo. The field containing the data category name is labelled "Data category name" for the canonical name, and is labelled by the language name for language-specific values (for instance, French). In the following example the canonical name is "partOfSpeech", the English name is "part of speech", and the French name is "catégorie grammaticale". The English name and the canonical name are often the same or similar, as English is the working language of the DCR. Data category name: partOfSpeech English: part of speech French: catégorie grammaticale
Source	Free text	Source of the data category name