



# SLOVENSKI STANDARD

## oSIST ISO/DIS 12620-1:2021

01-december-2021

---

### Upravljanje terminoloških virov - Podatkovne kategorije - 1. del: Specifikacije

Management of terminology resources -- Data categories - Part 1: Specifications

Gestion des ressources terminologiques -- Catégories de données - Partie 1: Spécifications

**(standards.iteh.ai)**

Ta slovenski standard je istoveten z: **ISO/DIS 12620-1**

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

#### ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

**oSIST ISO/DIS 12620-1:2021**

**en**

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

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

# DRAFT INTERNATIONAL STANDARD

## ISO/DIS 12620-1

ISO/TC 37/SC 3

Secretariat: DIN

Voting begins on:  
2021-09-30Voting terminates on:  
2021-12-23

---

---

## Management of terminology resources — Data categories —

### Part 1: Specifications

*Gestion des ressources terminologiques — Catégories de données —  
Partie 1: Spécifications*

ICS: 35.240.30; 01.020

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

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.



Reference number  
ISO/DIS 12620-1:2021(E)

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

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>



### **COPYRIGHT PROTECTED DOCUMENT**

© ISO 2021

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

Published in Switzerland

# Contents

	Page
Foreword.....	iv
Introduction.....	v
<b>1 Scope.....</b>	<b>1</b>
<b>2 Normative references.....</b>	<b>1</b>
<b>3 Terms and definitions.....</b>	<b>1</b>
<b>4 Data categories and data category specifications.....</b>	<b>3</b>
<b>5 General recommendations for data category specifications.....</b>	<b>4</b>
<b>6 Detailed requirements for documenting a data category.....</b>	<b>5</b>
6.1 Identifiers and names.....	5
6.1.1 A unique and stable mnemonic identifier.....	5
6.1.2 A unique and persistent identifier (PID).....	5
6.1.3 A unique canonical data category name.....	5
6.1.4 Language-specific data category names.....	5
6.2 Conceptual domains and data category types.....	5
6.3 Data elementarity.....	6
6.4 Profiles.....	6
<b>7 Referencing data categories.....</b>	<b>6</b>
<b>8 Harmonizing and vetting data categories.....</b>	<b>7</b>
<b>9 Management.....</b>	<b>8</b>
<b>Annex A (informative) Structure of a Data Category Specification.....</b>	<b>9</b>
<b>Bibliography.....</b>	<b>12</b>

oSIST ISO/DIS 12620-1:2021  
<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

## ISO/DIS 12620-1:2021(E)

### Foreword

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

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

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

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

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

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

This first edition of ISO 12620-1, together with ISO 12620-2<sup>1)</sup>, cancels and replaces ISO 12620:2019, which has been divided into parts and technically revised.

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

ISO 12620:2019, *Language and terminology — Data category specifications*, 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. The current version of the standard ISO 12620-1 has been narrowed to focus on the structure and rationale associated with data category specifications per se. Those sections of the standard that deal with the creation and maintenance of Data Category Repositories have been moved to ISO 12620-2, *Management of terminology resources — Data categories — Part 2: Repositories*.

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](http://www.iso.org/members.html).

1) Under preparation. (Stage at the time of publication ISO/DIS 12620-2.)

## 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 terminology resources, lexical 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.

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

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

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

[oSIST ISO/DIS 12620-1:2021](https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021)

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>



# 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. 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 12620-2,<sup>2)</sup> *Management of terminology resources — Data categories — Part 2: Repositories*

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

<https://standards.iteh.ai/catalog/standards/sist/23c21a53-e6d5-40a2-a89b-2bb806de60b2/osist-iso-dis-12620-1-2021>

### 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/ could 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/.

2) Under preparation. (Stage at the time of publication ISO/DIS 12620-2.)

## ISO/DIS 12620-1:2021(E)

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/ could, 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 can be implemented as a simple data category 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 are enclosed in forward slashes (e.g. /part of speech/).

[SOURCE: ISO 30042:2019, 3.8, modified – admitted term “DC” 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* (3.9).

## 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, 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 repository of *data category specifications* (3.5)

EXAMPLE DatCatInfo, a DCR for language resources (see Reference [5])

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

**3.7****data category selection**

DC selection

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 used within a research discipline or a specific application or project.

**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 <https://datcatinfo.termweb.eu/datcat/DC70>

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

**3.9****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: 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 picklist 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 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.