INTERNATIONAL STANDARD

ISO 12620

Second edition
2009-12-15

Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources

Terminologie et autres ressources langagières et ressources de contenu — Spécification de catégories de données et gestion d'un registre de catégories de données pour les ressources langagières

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

Reference number
ISO 12620:2009(E)

© ISO 2009

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

COPYRIGHT PROTECTED DOCUMENT

© ISO 2009

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

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

 © ISO 2009 – All rights reserved

Contents

Page

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2bc2-4ff1-add6-3f19085dfc2a/iso-12620-2009

© ISO 2009 – All rights reserved

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.

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

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

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.

ISO 12620 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 3, *Systems to manage terminology, knowledge and content*.

This second edition cancels and replaces the first edition (ISO 12620:1999), which has been technically revised.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

© ISO 2009 – All rights reserved

Introduction

Data associated with language resources are identified, collected, managed, and stored in a wide variety of environments. Data items appearing in individual language resources are themselves referred to in this International Standard as *data categories*, a designation commonly used in the environment of ISO Technical Committee ISO/TC 37. *Data categories* as cited in ISO/TC 37 standards correspond to *data element concepts* in the ISO/IEC 11179 series of standards, but differ slightly in terms of values defined. Differences in approach among different kinds of language resources and individual system objectives 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 thematic domain (for example, among terminological resources, lexicographic resources, annotated text corpora, etc.), at least at the interchange level, contributes to system coherence and enhances the re-usability of data. Procedures for defining data categories in a given thematic domain also need to be uniform in order to ensure interoperability among data categories, which becomes problematic if they are defined in individual data category registries.

The creation of a single global Data Category Registry (DCR) for all types of language resources treated within the ISO/TC 37 environment provides a unified view of the various applications of such a reference resource. This universal registry is designed to facilitate a wide range of Data Category Selections (DCS) needed in conjunction with all current or future standardization projects. ISO/TC 37 or any of its sub-committees can resolve at any time to designate specific *thematic domains* to deal with the management of those DCSs. The following thematic domains, among others, have been recognized as definable subsets of the DCR for language resources:

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/655b8151-2be2-48f4-ad46-319683df2e27/iso-12620-2009

— “Terminology”: ISO 16642:2003 explicitly refers to a set of reference data categories for terminology representation. Some of the data categories include general-purpose data management categories (for example, */source/*[1], */responsibility/*, */date/*, etc.) as well as linguistically oriented ones (for example, */partOfSpeech/*). Many of these data categories are relevant to a variety of different language resources, not just to terminology management, and form the core of the DCR as described in this International Standard;

— “Semantic Content Representation” and “Lexical Resources”: the DCR serves as a reference for the descriptors that are used throughout ISO/TC 37-related language resources, for instance, in terminology management systems, at various levels of linguistic annotation (for example, morphosyntactic, syntactic, and discourse levels), for lexical representation [natural language processing (NLP) lexicons, including machine translation (MT) dictionaries, etc.], or for specific applications such as metadata for language resources, query languages or multilingual data representation (for example, translation memories, localization files, etc.);

— “Language Codes”: ISO 639-1 and ISO 639-2 contain codes for approximately 650 languages. ISO 639-3 extends this number by an order of magnitude, with a clearer separation between the description of the language and its coding proper [1][2][3]. Including the reference set of language identifiers in the DCR in response to the evolution of the ISO 639 family of standards provides an essential element of any linguistic annotation or representation scheme.

— “Lexicography”: the deployment of the DCR will include data categories for the description of lexicographic data as cited in ISO 1951:2007[4] in order to ensure that the formats used for describing lexicographical (SC 2), terminological (SC 3) and NLP-oriented (SC 4) data are comparable and compatible.

---

1) Names that function as class names are capitalized. When a name functions as an attribute, it is represented in bold face with the **+** convention: i.e. **+administration record** as opposed to Administration Record. This function is context-dependent. Terminal values used with the data model appear in normal face bracketed by hyphens: -standardized name-. Data category names are represented using the convention */source/*. Data categories are themselves defined in the DCR, not in this International Standard.

© ISO 2009 – All rights reserved

The DCR will eventually contain all ISO/TC 37 data categories, with their complete history, data category descriptions, and attendant metadata. It is not, however, the intent of this International Standard to define an ontology of language resources within ISO/TC 37. Nevertheless, the definition of the DCR has avoided any choices that would hamper further work in this direction.

This document is intended to provide a background in the context of ISO/TC 37 on the various issues that have to be considered in order to implement a global DCR that can be used for the full range of language resources. More precisely, this document addresses the following issues:

— the role of data categories for use with language resources;

— requirements that have been identified with respect to information content and overall management;

— a description of the organization of the DCR;

— an interchange format for data categories, the DCIF (Data Category Interchange Format);

— management procedures for the DCR.

Specific user-oriented instructions and procedures pertaining to the implementation and use of the DCR are available on-line at http://www.isocat.org.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

 © ISO 2009 – All rights reserved

Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources

1 Scope

This International Standard provides guidelines concerning constraints related to the implementation of a Data Category Registry (DCR) applicable to all types of language resources, for example, terminological, lexicographical, corpus-based, machine translation, etc. It specifies mechanisms for creating, selecting and maintaining data categories, as well as an interchange format for representing them.

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.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 8601:2004, *Data elements and interchange formats — Information interchange — Representation of dates and times*

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/d25b8151-2be2-4f7b-add6-319683dfc2a7/iso-12620-2009

ISO/IEC 11179-1:2004, *Information technology — Metadata registries (MDR) — Part 1: Framework*

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

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 11179-1:2004 and the following apply. Terms and definitions have evolved in the terminology community, represented here by citations from ISO 1087-2, independently of the terminology of the metadata community, which results in slightly different and at times overlapping concepts in the two communities of practice.

3.1 Data elements and data categories

3.1.1
data element
⟨language resources⟩ unit of data that, in a certain context, is considered indivisible

[ISO 1087-2:2000, 6.11]

NOTE In terminology work, an individual field, for example, /*term*/, in a single terminological entry has been viewed as a data element and an instantiation of a **data category** (3.1.3).

© ISO 2009 – All rights reserved

3.1.2
data element
DE
⟨metadata standards⟩ unit of data for which the definition, identification, representation and value domain are specified by means of a set of attributes

[ISO/IEC 11179-1:2004, 3.3.8]

3.1.3
data category
DC
result of the specification of a given data field

[ISO 1087-2:2000, 6.14]

EXAMPLE */partOfSpeech/, /grammaticalGender/, /grammaticalNumber/;* the values associated with these items (for example, */noun/, /verb/, /feminine/, /plural/,* etc.) are also data categories according to this International Standard, but values of this type are not viewed as **data element concepts** (3.1.4) in the ISO/IEC 11179 family of standards.

NOTE 1 A data category is an elementary descriptor in a linguistic structure or an **annotation scheme** (3.1.15).

NOTE 2 A data category corresponds closely, but not identically, to a data element concept in ISO/IEC 11179.

NOTE 3 In running text, such as in this International Standard, **data category** (3.1.3) names are set off using forward slashes and italics. In some implementations, camel case is used instead of using white space between words in the data category name.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.1.4
data element concept
concept for which the definition, identification and **conceptual domain** (3.1.5) are specified independently of any particular representation

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

[ISO/IEC 11179-1:2004, 3.3.9]

3.1.5
conceptual domain
set of valid value meanings

NOTE 1 Adapted from ISO/IEC 11179-1:2004, 3.3.6.

NOTE 2 The value meanings in a conceptual domain may be enumerated, further specified by additional constraints or expressed via a description. For instance, the **data category** (3.1.3) */term/* is described by its definition and thus constrained from properly containing, for example, contextual information or grammatical information, but it would be impossible to enumerate all values associated with this data category.

3.1.6
value domain
set of permissible values

[ISO/IEC 11179-1:2004, 3.3.38]

3.1.7
complex data category
data category (3.1.3) that has a **conceptual domain** (3.1.5)

3.1.8
open data category
complex data category (3.1.7) whose **conceptual domain** (3.1.5) is not restricted to an enumerated set of values

© ISO 2009 – All rights reserved

3.1.9
open conceptual domain
conceptual domain (3.1.5) associated with an **open data category** (3.1.8)

3.1.10
constrained data category
complex data category (3.1.7) whose **conceptual domain** (3.1.5) is non-enumerated, but is restricted to a constraint specified in a schema-specific language or languages

3.1.11
constrained conceptual domain
conceptual domain (3.1.5) associated with a **constrained data category** (3.1.10)

3.1.12
simple data category
data category (3.1.3) with no **conceptual domain** (3.1.5)

3.1.13
closed data category
complex data category (3.1.7) whose **conceptual domain** (3.1.5) is restricted to a set of enumerated **simple data categories** (3.1.12) making up its **value domain** (3.1.6)

3.1.14
closed conceptual domain
conceptual domain (3.1.5) associated with a **closed data category** (3.1.13)

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.1.15
annotation scheme
set of descriptors together with their syntax, semantics and condition of use, intended to provide descriptive or interpretive information relevant to a language resource

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

NOTE TEI ODD (One Document Does it All) is an example of an annotation scheme.[10]

3.2 Data Category Registry

3.2.1
Data Category Registry
DCR
set of **data categories** (3.1.3) to be used as a reference for the definition of linguistic **annotation schemes** (3.1.15) or any other formats used in the area of language resources

3.2.2
data category specification
set of attributes used to fully describe a given **data element concept** (3.1.4)

NOTE The abbreviation “DCS” is associated with the **Data Category Selection** (3.2.3) and should not be confused with the data category specification.

3.2.3
Data Category Selection
DCS
set of **data categories** (3.1.3) selected from the **DCR** (3.2.1)

NOTE 1 A DCS can represent the data categories used within a **thematic domain** (3.4.3) or a selection of data categories used for a specific application or project. In the latter case, the DCS may draw data categories from more than one thematic domain.

NOTE 2 A DCS can be expressed as a simple list of data categories, or it can be output in a form that contains the entire content of their associated **data category specifications** (3.2.2), thus incorporating the full set of constraints associated with the DCS. It can also be expressed using a schema notation such as W3C XML Schema[11] or Relax NG[12], which also comprises the list of data categories together with their associated constraints.

© ISO 2009 – All rights reserved

3.3 Data category specification components

3.3.1
DCR data model
logical representation of data structure and dependencies in the **DCR** (3.2.1)

NOTE 1 The DCR data model is represented as a UML class diagram[13].

NOTE 2 The definition above is based on ISO/IEC 11179-1:2004, 3.2.7, where “data model” is defined as a “graphical and/or lexical representation of data, specifying their properties, structure and inter-relationships”.

3.3.2
Global Information
GI
technical and administration information applying to the entire data collection

[ISO 16642:2003, definition 3.7]

EXAMPLE Title of the data collection, revision history.

3.3.3
administration information section
class in a **data category specification** (3.2.2) pertaining to the submission, registration, balloting and approval of data category specifications submitted to and maintained in the **DCR** (3.2.1)

3.3.4
registration group
class associated with the **administration information section** (3.3.3) that provides information related to the **Registration Authority (RA)** (3.4.2) responsible for the administered item

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.3.5
submission group
class associated with the **administration information section** (3.3.3) that provides information on persons or groups responsible for submitting the administered item

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-5f968a3dfc2a/iso-12620-2009

3.3.6
decision group
class associated with the **administration information section** (3.3.3) that provides information on the review and balloting process associated with the administered item

3.3.7
stewardship group
class associated with the **administration information section** (3.3.3) that provides information on the individual or group responsible for the maintenance of the administered item

3.3.8
description section
class pertaining to the **data category** (3.1.3) name and the **data element concept** (3.1.4) documented in a **data category specification** (3.2.2)

NOTE Definitions, explanations, and notes comprise some of the kinds of information included in the description class of a data category specification.

3.3.9
data element name
class in a **data category specification** (3.2.2) that lists and categorizes permissible names that can be associated with the **data category** (3.1.3)

© ISO 2009 – All rights reserved

3.3.10
language section
class in a **data category specification** (3.2.2) that provides **working language** (3.3.12) equivalents for **data category** (3.1.3) names and other descriptive information included in a **data category specification** (3.2.2)

3.3.11
linguistic section
class in a **data category specification** (3.2.2) that constrains the **conceptual domain** (3.1.5) for a given **object language** (3.3.13)

3.3.12
working language
language used to describe objects

[ISO 16642:2003, 3.21]

3.3.13
object language
language being described

[ISO 16642:2003, 3.10]

3.3.14
name section
class in the **language section** (3.3.10) that lists variant names for the **data category** (3.1.3) treated in a **data category specification** (3.2.2)

iTeh STANDARD PREVIEW
(standards.iteh.ai)

NOTE Variant names can be equivalents in other languages or names that may be used in the same language but in related disciplines or working environments.

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

3.4 DCR management

3.4.1
Data Category Registry Board
DCR Board
DCRB
group of **experts** (3.5.3) designated by the participating (P) members of the Technical Committee whose duty it is to ensure that the scope and the coherence of the **Data Category Registry** (3.2.1) are maintained

NOTE The DCRB has the status of a Validation Team (VT) according to Annex ST of the ISO Supplement to the ISO/IEC Directives.[9]

3.4.2
Registration Authority
RA
organization authorized to register data items and/or other information objects and to maintain them in a repository

NOTE Typically these kinds of information objects can comprise codes, such as the language codes defined in the ISO 639 family of standards, data categories, and other public identifiers. Registration Authorities are governed by International Standards, but the repositories themselves are not generally included in published standards.

3.4.3
thematic domain
class of applications identified by the similarity of the data structures they need to manipulate

EXAMPLES Terminology, lexicography, morphosyntactic annotation.

© ISO 2009 – All rights reserved

3.4.4
thematic domain group
TDG
group of **experts** (3.5.3) in charge of selecting and maintaining the **data categories** (3.1.3) that are relevant for a **thematic domain** (3.4.3)

NOTE Thematic domain groups have the status of the Maintenance Team (MT) according to Annex ST of the ISO Supplement to the ISO/IEC Directives.[9]

3.4.5
thematic domain profile
profile
representation within a **data category specification** (3.2.2) of the **thematic domain** (3.4.3) with which a **data category** (3.1.3) is associated

NOTE A data category may have several thematic domain profiles, indicating that it is used by several thematic domains. Until a data category specification is assigned to a TDG, the profile value is set to -private-.

3.5 Roles

3.5.1
chair of the Data Category Registry Board
chair of the DCR Board
chair of the DCRB
individual appointed by the ISO/TC 37 plenary who has the responsibility of administering the work of the **DCRB** (3.4.1)

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.5.2
chair of a thematic domain group
TDG chair
individual appointed by the ISO/TC 37 sub-committee associated with a **TDG** (3.4.4) who has the responsibility for administering the work of the TDG

ISO 12620:2009
https://standards.iteh.ai/catalog/standards/sist/625b8151-2be2-4ff4-add6-319683dfc2a7/iso-12620-2009

3.5.3
expert
individual with special knowledge, skill, or other interest who registers to participate in the work of the **DCR** (3.2.1)

3.5.4
judge
expert appointed by the **chair of a thematic domain group** (3.5.2) to participate in the approval process for any given **data category specification** (3.2.2) or **Data Category Selection** (3.2.3) submitted for standardization

3.6 Data exchange

3.6.1
Data Category Interchange Format
DCIF
export format for **data categories** (3.1.3) grouped as a **Data Category Selection** (3.2.3) designed to facilitate their usability in external applications

3.6.2
snapshot
capture of the status of a data resource at a given moment in time

NOTE Data resources are frequently archived in the form of snapshots, which can then be identified as versions of the resource.

© ISO 2009 – All rights reserved

3.6.3
persistent identifier
PID
unique Uniform Resource Identifier (URI) that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership

4 Role of data categories in language resource management

4.1 Overview

Data category specifications identify the individual information units making up a data collection or annotation scheme for a given language resource. A data category specification provides the formal representation of a data category, which shall comprise the specific attributes that document it (for example, the data category name, definition, examples, comments, etc.). It shall also provide the context for its creation and management within the DCR. Groups of data categories subsetted from the global set making up the DCR comprise Data Category Selections (DCS). As specified in ISO 16642, *Terminological Markup Framework (TMF)*, a DCS shall define, in combination with a data model, the various constraints that apply to given thematic-domain or application-specific information structures or interchange formats.

Figure 1 shows possible uses for a DCS. Depending on the application involved, such a DCS may be merely a list of data categories that points back to the complete specifications in the DCR, or it can be represented by a complete subset or even superset of the DCR, comprised of such a list plus the definitions and constraints associated with the individual data category specifications.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO 12620:2009

From a wider perspective, a formal model for representing data categories shall account for the fact that apart from pure computer use, a data category specification can be intended for human use as well. For instance, such specifications can form the core of a DCS, which can be published either as a paper document, made available as an electronic resource, or identified as a subset of the ISO/TC 37 DCR. Typically, the designers of a given markup language or data management system will query the DCR in order to create their individual application-specific DCSs by selecting a subset of data categories from the global DCR. Finally, providing a precise description of the data categories used within a given data collection in reference to the DCR allows for a quick diagnosis of the compatibility of this collection with any other particular computer application and thus can act as metadata for this collection.

Figure 1 also presents the notion of a DCS, for example, the choice of a specific set of data categories taken from the global DCR for use in a given thematic domain within the framework of language resources, or in a specific application. The diagram exemplifies the various roles of a DCS in the process of defining and using any linguistic annotation scheme. Viewed from this perspective, a DCS is primarily intended to contribute to the specification of the DCR annotation scheme in combination with the data model that expresses the general organization of the DCR. This kind of selection guarantees a certain degree of interoperability between two or more data structures by facilitating the comparison of the selected data categories as well as the constraints imposed on them, for instance the nodes in the DCR data model that correspond to the positions where each category is allowed to occur in the individual data structures, such as in specific applications or annotation schemes. In this scenario, the DCS for each of the structures in question can be expressed as, for example, a Relax NG[12] or W3C XML Schema[11], and XSL filters can be used to output relevant data from the Data Category Interchange Format (DCIF) in alternative formats (see 7.9).

In addition, the DCS can be seen as a documentary source for the linguistic annotation scheme in question. Indeed, because it contains the list of all data categories that can be used in conjunction with the annotation scheme, it is probably the best source of information for potential users or implementers who want to know whether a given data category corresponds to their needs.

Furthermore, the DCS can be attached (or referenced) in any data transmission process to provide the receiver with all the information needed to interpret the content of the information being transmitted. In particular, this procedure should allow linguistic data expressed in various kinds of XML representations to be sent or received in the most transparent way.

© ISO 2009 – All rights reserved