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

Reference number
ISO 12620:2009(E)

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

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) 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.

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:

— "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.

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.

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.

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

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).

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.

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/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

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)

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

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.

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

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

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)

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)

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.

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.

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)

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

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.

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.

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.

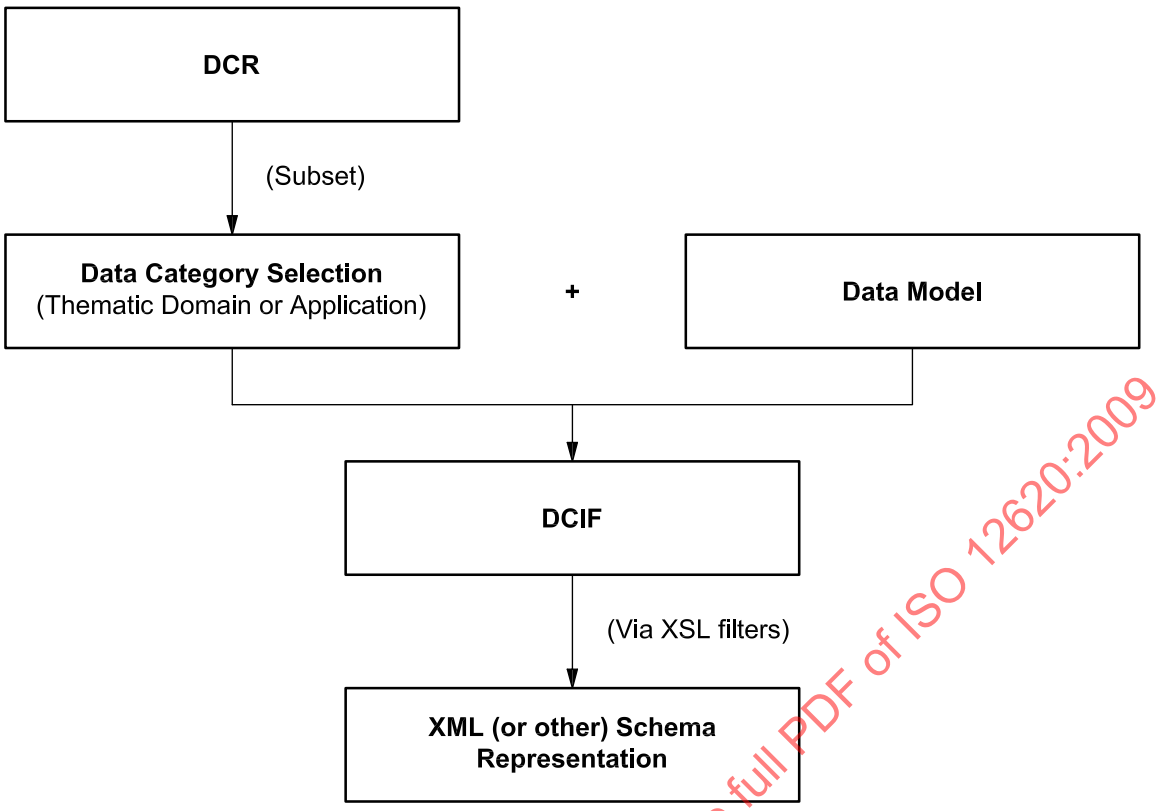


Figure 1 — Role of Data Category Selections in the context of the definition of linguistic annotation schemes

4.2 Variety of Data Category Selections (DCSs)

Figure 2 illustrates the relationship between individual data category specifications, the DCR, and any one of the possible DCSs that represents a subset of the DCR. The largest circle represents the entire collection of data category specifications included in the DCR, and the smaller internal circles represent DCSs which are subsets of the DCR. The very small, individual background patterns in the drawing correspond to individual data category specifications, each describing a given data category concept using the data category attributes set down for the DCR with reference to the attributes defined in ISO/IEC 11179. For instance, one small pattern element might represent the data category specification for the */term/* data category.

Some data categories included in the DCR are pertinent to a single thematic domain within the broader field of terminology and other language and content resources. For instance, a */conceptIdentifier/* is probably unique to terminological resources (although not prescriptively), or a */senseNumber/* is probably specific to lexicographical resources. On the other hand, many data categories, frequently those of a strictly linguistic nature such as */partOfSpeech/, /grammaticalGender/, /grammaticalNumber/,* etc., are common to a wide variety of resources. To be sure, these categories may not always have the same function in different thematic domains, but they nevertheless represent the same essential concept in different kinds of resources. Hence each thematic domain shall contribute all its data categories in the form of data category specifications to the global DCR, while at the same time identifying those data categories that it shares with other kinds of resources. The subset of data categories and their specifications used in a thematic domain shall comprise a domain-specific DCS taken from the DCR.

As noted, the oval shapes in the diagram represent such DCS subsets. A further subset can be selected from one or more thematic domain DCSs for use in a given application or collaborative environment. The octagon shown in Figure 2 represents an example of an application-specific subset. This particular application is wholly contained within the DCS for terminological data categories, so it is apparently designed for use with a

terminological application, although some of the data category specifications contained in this subset are common to several different kinds of language resources. It should thus be noted in this context that some applications can indeed draw data categories from several thematic domain DCSs. Theoretically, it is also possible for a DCS to be in part a superset if it happens to include categories that are not currently part of the DCR. In such cases, it is recommended that designers or users of such a set register their new data categories in the DCR.

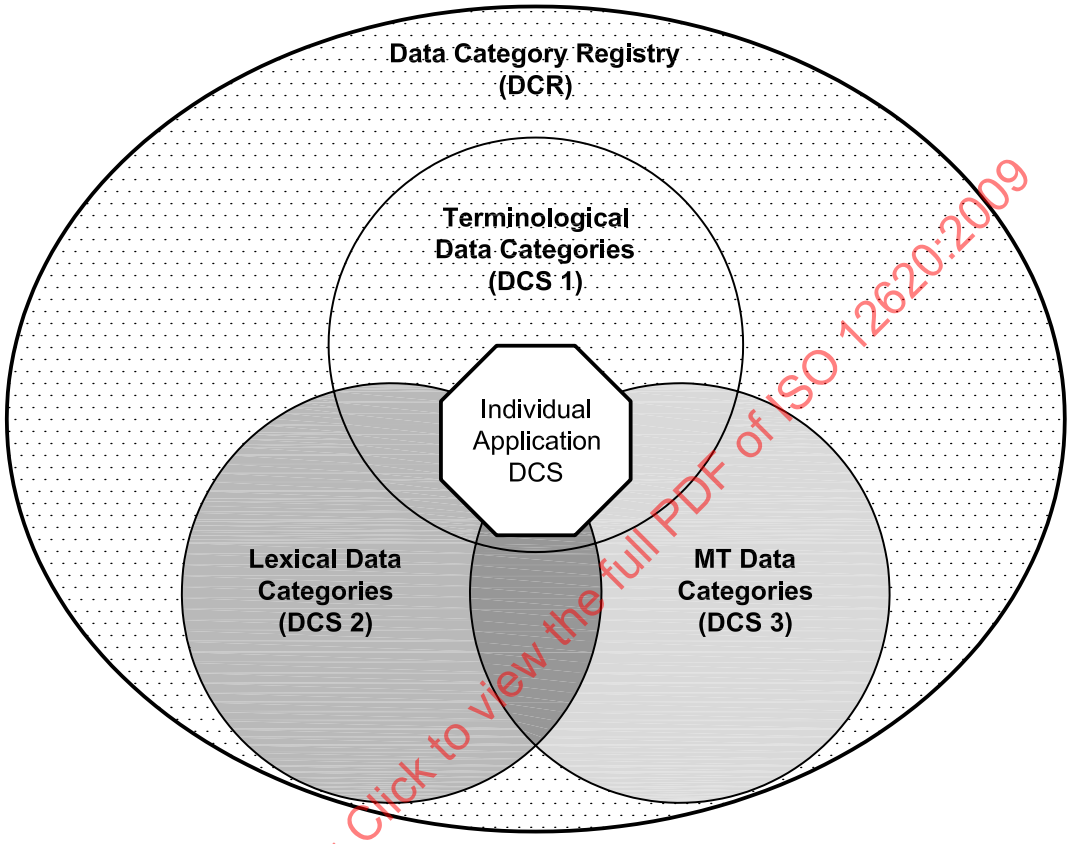


Figure 2 — Individual application-oriented DCS viewed in the context of various TDG-oriented DCSs

5 Requirements for the implementation of a DCR for language resources

This clause outlines the basic requirements that a DCR should fulfil in order to support standardization activities within ISO/TC 37.

The ISO/TC 37 DCR shall:

— Provide a reference repository for data categories and related information for all the existing or future standards in ISO/TC 37 that involve data modelling or data interchange.

— Be available online, free of charge.

— Register existing practices by associating each data category with the way it is implemented in specific projects or initiatives. This may consist in registering various types of encodings, from basic codes (for example, 'f' for feminine in EAGLES [15] morphosyntactic descriptions) to actual XML representations.

— Provide names and definitions in a variety of languages.

— Describe the use of each data category in a variety of object and working languages. This may consist of a specific definition (for instance when the data category has a slightly different application scope), some

usage notes, examples, or list of values (for example, the value domain of */gender/* is {*/masculine/*, */feminine/*} in French, and {*/masculine/*, */feminine/*, */neuter/*} in German).

— Describe the use of data categories in a variety of resources, where necessary.

— Associate administration information with each data category so that it is possible to trace the submission, acceptance or revision of the data category.

— Associate each data category with one or several profile values corresponding to the thematic domains where the data category is relevant (for instance, */partOfSpeech/* is relevant for both terminology and lexicography).

— Provide a mechanism by which a thematic domain group in ISO/TC 37 can submit a group of data categories relevant to its scope of activities.

— Provide private workspaces where individuals and working groups outside ISO/TC 37 can generate or upload their own data category specifications and selections, to publish and share them, and, when desired, to submit data categories for registration and approval.

— Be updated on a regular basis by integrating, according to defined rules, proposals from experts in the field.

— Be implemented in accordance with the main principles stated in the ISO/IEC 11179 family of standards.

— Provide stable worldwide accessibility by maintaining distributed mirror sites.

— Provide for persistent identifiers by introducing a system of stable references for individual data specifications.

— Maintain secure, best practices for archiving procedures.

— Provide ISO with periodic data snapshots of those data categories that have been approved following procedures spelled out in Annex ST of the ISO Supplement to the ISO/IEC Directives and Clause 8 of this International Standard, with the understanding that ISO shall make this subset of the DCR available through its standard practices without compromising the availability of these items within the environment of the DCR itself.

6 Registration Authority for the ISO/TC 37 DCR

The ISO/TC 37 DCR shall be implemented under the auspices of a Registration Authority (RA), as defined in ISO/IEC Directives, which has the duty to register data categories in accordance with the rules defined in this International Standard.

The Technical Management Board has appointed the Max Planck Institute for Psycholinguistics[2)] to be the Registration Authority for the ISO/TC 37 DCR.

The RA shall maintain the DCR as a web service at http://www.isocat.org. All user functions and support shall be available through this website.

2) At the time of publication of this International Standard, the ISO/TC 37 DCR Registration Authority is held by:

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

ISO maintains an online listing of maintenance and registration authorities at http://www.iso.org/rama. Users are encouraged to consult this listing for the most up-to-date information concerning maintenance agencies and registration authorities.

7 Representation of data categories used in language resources

7.1 Introduction

This section describes the data model of the DCR for ISO/TC 37. The data model is specified in Unified Modelling Language (UML)[13], extended where necessary with additional constraints expressed in the Object Constraint Language (OCL)[16]. The complete data model is shown in an overview in Figure 3, broken out in three parts in Figures 4, 5 and 6, where the Data Category serves as the linking class for all figures.

The DCR shall consist of two main classes:

— Global Information (GI);

— one or more Data Category (DC) specifications.

Each data category specification shall consist of two mandatory classes:

— one dedicated to the administration and identification of the data category (the Administration Information Section);

— one dedicated to the documentation of the data category in one or more working languages and zero or more names as used in a given database, format or application (the Description Section).

Further enhancements may be made to the data category specification depending on the exact data category type. These include the following classes:

— one or more classes which describe the conceptual domain of the data category (the Conceptual Domain and subclasses);

— one or more classes which describe the conceptual domain and/or the use of the data category in the context of a specific object language (the Linguistic Section and subclasses).

Further classes associated with these main classes are described in more detail in the following subclauses.

7.2 Global Information class

The DCR shall make available the following items of information reflecting Global Information:

— the name and complete contact information of the Registration Authority (complete address, telephone and fax numbers, e-mail contact address);

— the date when the RA was appointed and other critical dates concerning versions of the RA;

— the name of the person primarily responsible for administration of the RA;

— historical information concerning previous RAs in the event that responsibility for the RA is transferred to another entity;

— short forms or abbreviations used in the DCR to reference the current RA and any previous RA(s) that may have been responsible for the registry;

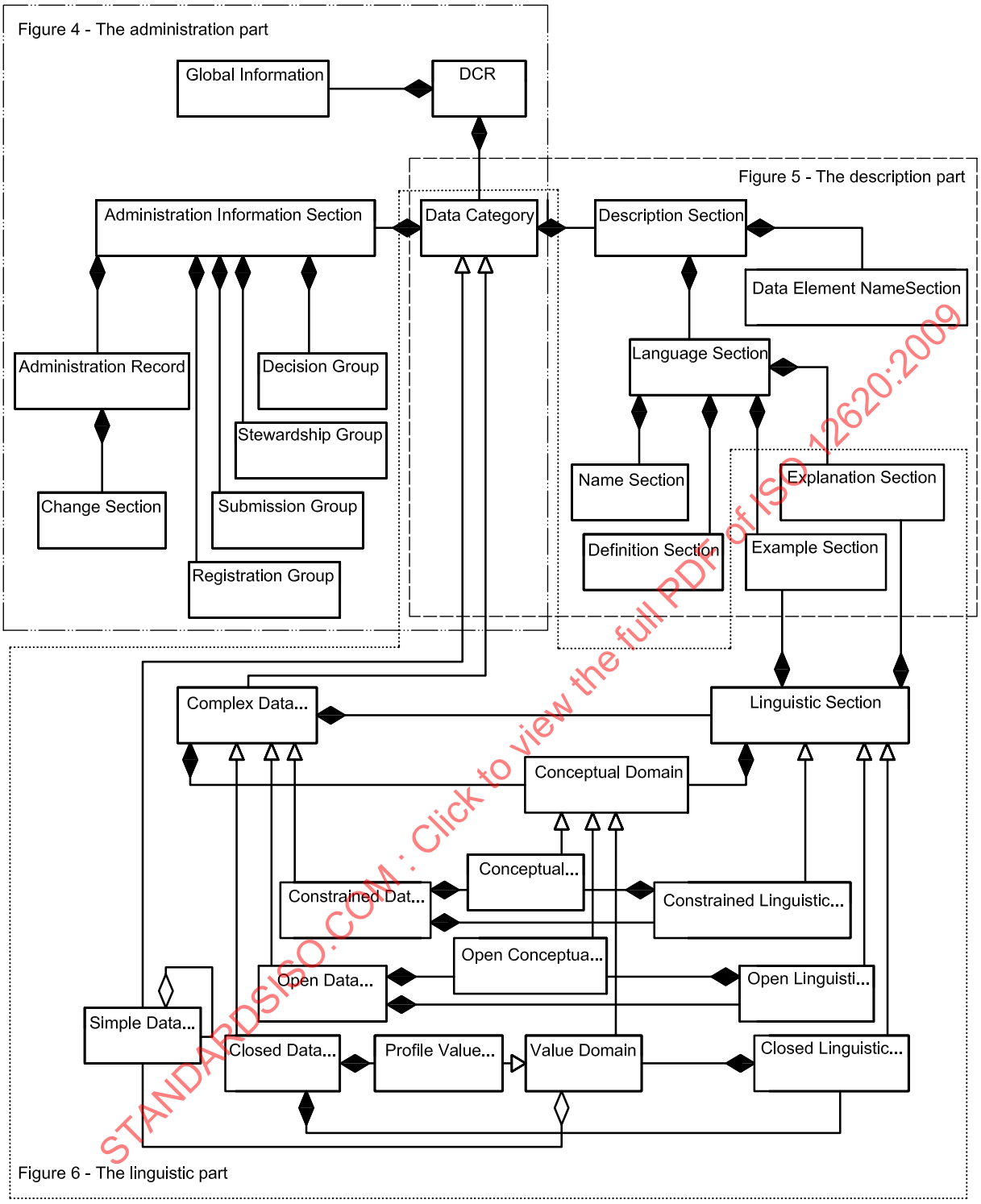
— names and affiliations of members of the DCRB;

— information on the RA agreement between ISO and the RA;

— the DCR mission statement;

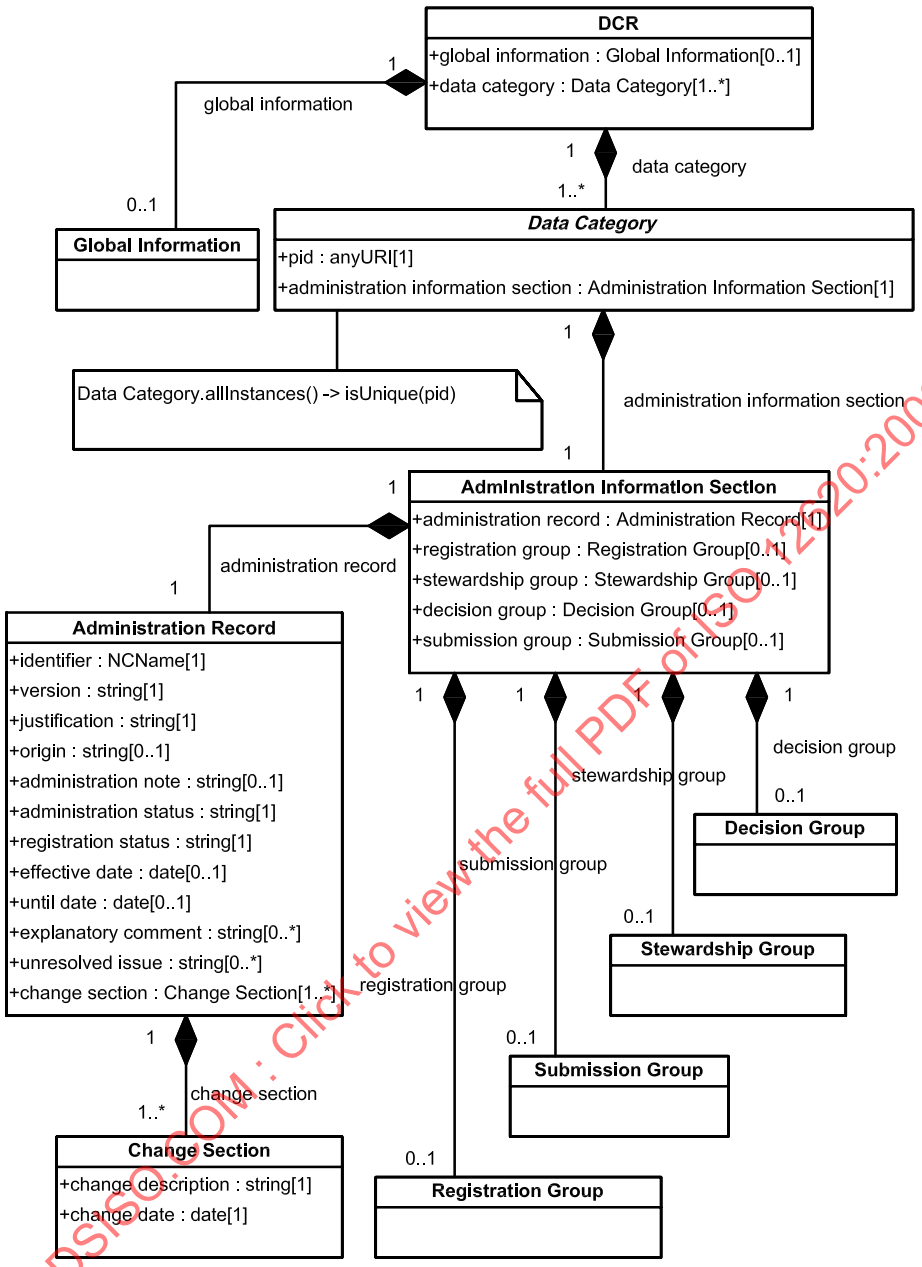
— a statement of legal responsibilities and restrictions.

This International Standard does not impose any additional specific constraints on this class. It is the responsibility of the RA and the DCRB to provide technical information as deemed fit.



△ inheritance relationship from subclass to superclass
♦ composition relationship between two classes
◊ aggregation relationship between two classes

Figure 3 — Overview of the DCR data model (see Figures 4, 5, and 6 for details)



△ inheritance relationship from subclass to superclass

♦ composition relationship between two classes

◊ aggregation relationship between two classes

Multiplicity values: 0..1 = 0 or 1; 0..* = 0 or more; 1 = 1 and only 1; 1..* = 1 or more instances may occur.

Figure 4 — Administration part of the DCR data model

7.3 Data Category classes

7.3.1 Data Category subtypes

Two primary subtypes of Data Category are maintained in the DCR and distinguished as follows:

a) Complex Data Categories, such as */term/* or */grammaticalGender/*;

b) Simple Data Categories, such as */masculine/, /feminine/,* etc.

These two basic types of data category are represented in the DCR model as two subclasses of Data Category. They are shown in Figure 6 in conjunction with the Linguistic Section (see 7.7).

Data Category contains the following attribute:

— **+pid** [1]: the persistent identifier (PID) of the data category; for standardized data categories, the RA shall endeavour to ensure that these identifiers will be resolvable, even when the data category is deprecated. When referencing a data category in a bibliographical or technical context, the **+pid** should be used, see also 7.8, Referencing a data category.

7.3.2 Complex Data Category

The Complex Data Category allows data categories to have a conceptual domain. The DCR data model supports three types of conceptual domains: open, constrained and closed. For specific object languages, the conceptual domain can be further restricted. Modelling of domains and restrictions is described in 7.6, Conceptual Domain classes and 7.7, Linguistic Section.

7.3.3 Simple Data Category

Simple data categories represent values associated with value domains of the complex data categories. In contrast to complex data categories, simple data categories can be part of a value hierarchy by participating in an 'is a' association. This feature allows users, for example, to declare a */properNoun/* to be a */noun/*. This type of association is not permitted between complex data categories because the description of more extensive concept hierarchies falls outside the scope of the DCR.

— **+is a** [0..1]: attribute used to point to the more generic simple data category with which a simple data category is associated.

This association is constrained in the sense that the simple data category and the more generic simple data category should share membership in at least one profile. Also the graph which is built by these associations has to be acyclic, in other words, a simple data category cannot be, directly or indirectly, its own more generic simple data category.

7.4 Administration Information Section

7.4.1 Associated classes

The Administration Information Section can be further decomposed into five associated classes:

— Administration Record, which groups together the information associated with the global management of the administered item.

— Registration Group, which contains information related to the RA as well as the DCRB.

— Submission Group, which contains information related to the entity that has submitted the data category to the registry. This entity may be either the thematic domain group that has selected the data category, and/or possibly the expert or expert group originating the submission.

— Stewardship Group, which contains the information identifying the thematic domain group responsible for the maintenance of the administered item, which acts as a Maintenance Team in the sense of Annex ST of the ISO Supplement to the ISO/IEC Directives [9].

— Decision Group, which contains the information associated with the decision process involved in the evaluation of the data category by the relevant thematic domain group and its validation by the DCRB.

7.4.2 Information to be expressed in the Administration Record Section

The Administration Record Section contains information on the identification and maintenance of the administered item. It is associated with the following attributes:

— **+identifier** [1; ISO/IEC 11179-3]: mnemonic string used to refer to the data category.

  According to ISO/IEC 11179-3, the identifier should be presented as an alphanumeric character string. For sake of legibility, the identifier may be based on a series of English words reflecting its actual meaning (for instance, */term/*, */normativeAuthorization/*, */preferredTerm/*), but such a practice should not preclude the use of additional names for the data category in English or any other language. Note also that these elements shall appear in camel case in the identifier.

  To enable the use of a data category identifier in XML vocabularies, it should be a valid local part of a qualified name, as defined for XML documents which conform to the XML namespaces recommendation[14].

  When referencing a data category in a bibliographical or technical context, the **+pid** should be used rather than the **+identifier**; see also 7.8, Referencing a data category.

— **+version** [1]: used to refine **+identifier** to indicate the version of the data category.

— **+administration note** [0..1; ISO/IEC 11179-3]: any general note about the administered item.

— **+administration status** [1]: a designation of the status in the administration process for handling registration requests under the stewardship of the DCRB. The following values may be used for the administration status:

  — -private-: the data category specification is restricted to an expert's private workspace or is shared in a private group, but has not been (and may never be) submitted to the standardization process;

  — -submission-: a data category specification has been submitted by an individual expert or group of experts (as documented in the Submission Group Section) to a given TDG (as documented in the Stewardship Group Section), thus initiating the DC selection and standardization process shown in Figure 9;

  — -pre-evaluation-: the chairs of the DCRB and the TDG have approved the proposal and advanced it to the evaluation stage within the thematic domain group;

  — -evaluation-: acceptance of the data category specification is under evaluation within the thematic domain group;

  — -rejected-TDG-: the data category specification has been rejected by the TDG;

  — -accepted-TDG-: acceptance of the data category specification by the TDG, as reflected in the Resolution of Acceptance shown in Figure 9;

  — -pre-validation-: preparation for validation by DCRB chair;

  — -validation-: the data category specification has been approved by the TDG, prepared by the TDG and DCRB chairs, and forwarded to the DCRB for final validation, that is, consideration and approval;

  — -accepted-: the data category specification has been validated and accepted by the DCRB for inclusion in the DCR;

  — -rejected-DCRB-: the data category specification has been rejected by the DCRB.

— **+registration status** [1; ISO/IEC 11179-3]: a designation of the status in the registration life-cycle of an administered item.

The following values may be used for **+registration status**, as excerpted from ISO/IEC 11179-6:2005:

— -candidate-: the data category specification has been proposed for progression through the DCR registration process;

NOTE The registration status of a data category specification is set to -candidate- until its administration status is finally confirmed as -accepted- (in which case the registration status becomes -standard-).

— -standard-: the DCRB has confirmed that the data category specification is of sufficient quality and of broad interest for use in the DCR community;

— -deprecated-: the DCRB has confirmed that the data category specification is not or is no longer recommended for use in the registry community;

— -superseded-: the DCRB has confirmed that the data category specification is no longer recommended for use in the DCR community and has designated a successor data category specification as preferred for use.

— **+effective date** [0..1]: the date a data category specification has/will become available to DCR users, which shall be stated in the form YYYY-MM-DD as per ISO 8601:2004 and ISO/IEC 11179-3.

— **+change section** [1..*]: information on when the data category specification most recently underwent a change, including information on when the data category specification is initially created (for instance in an expert's private workspace); the DCR shall record all changes.

— **+explanatory comment** [0..*; ISO/IEC 11179-3]: descriptive comments about the data category specification.

— **+origin** [0..1; ISO/IEC 11179-3]: source (document, project, discipline or model) for the data category specification.

— **+justification** [1]: a short description justifying why the data category should be included in the registry.

— **+unresolved** issue [0..*; ISO/IEC 11179-3]: a problem that remains unresolved regarding proper documentation of the data category specification.

— **+until date** [0..1; ISO/IEC 11179-3]: the date (which shall be in the form YYYY-MM-DD as specified in ISO 8601:2004) a data category specification is no longer effective in the registry. This information is set when the registration status of the data category specification changes to -deprecated- or -superseded-.

The Change Section class, used by **+change section**, records the following information:

— **+change date** [1]: the date (which shall be YYYY-MM-DD as specified in ISO 8601:2004) a modification was made in the data category specification.

— **+change description** [1]: free text description of the modification made in the data category specification (for instance, “definition updated…”).

7.4.3 Information to be expressed in the Registration Group class

No specific constraint is imposed by this International Standard on this component. It is left to the Registration Authority to provide a technical note that explains the implementation of this component when it is actually used.

7.4.4 Information to be expressed in the Submission Group class

No specific constraint is imposed by this International Standard on this component. It is left to the Registration Authority to provide a technical note that explains the implementation of this component when it is actually used.

7.4.5 Information to be expressed in the Stewardship Group class

No specific constraint is imposed by this International Standard on this component. It is left to the Registration Authority to provide a technical note that explains the implementation of this component when it is actually used.

7.4.6 Information to be expressed in the Decision Group class

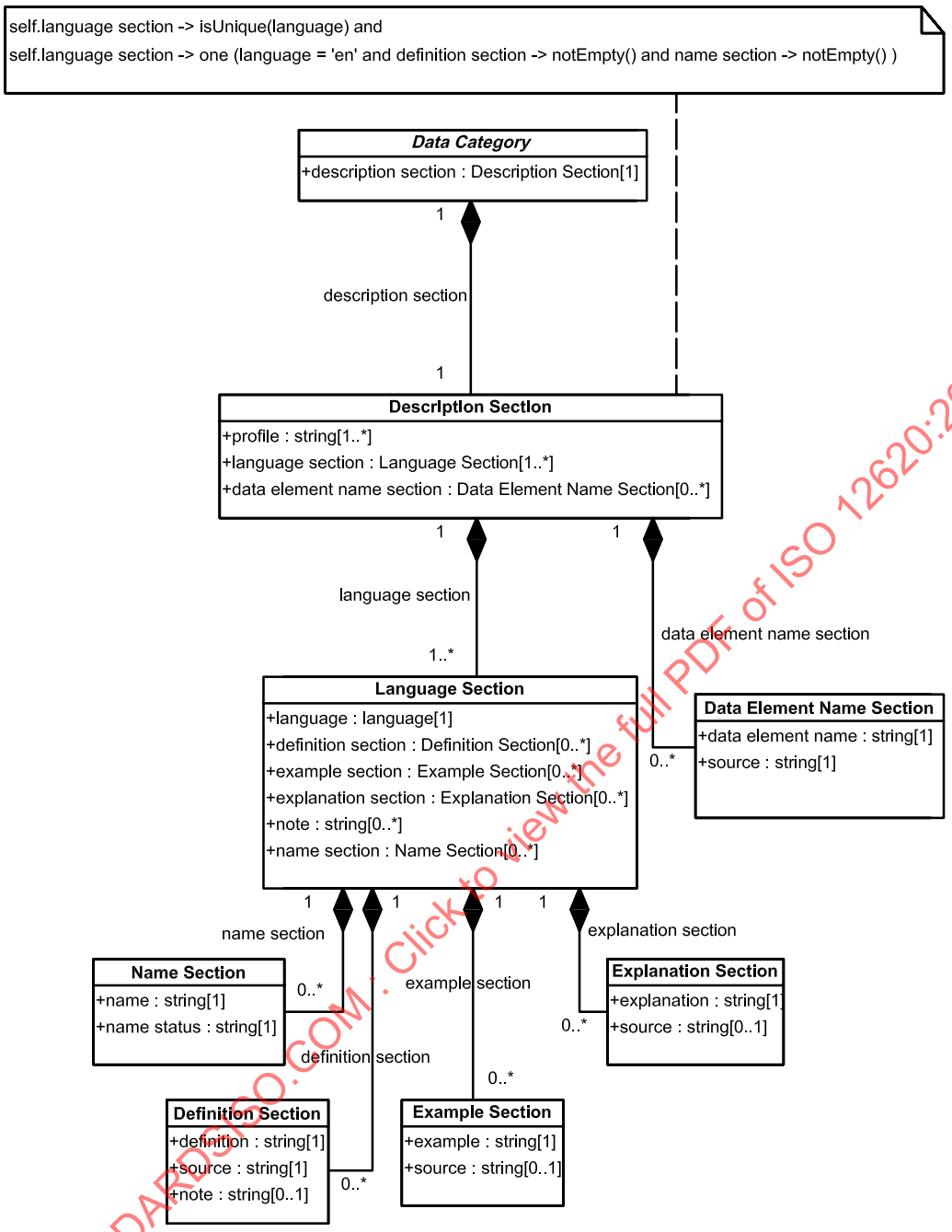
No specific constraint is imposed by this International Standard on this component. It is left to the Registration Authority to provide a technical note that explains the implementation of this component when it is actually used.

7.5 Documenting data categories

7.5.1 Information to be expressed in the Description Section

The description part of a data category specification can be viewed as having one or more Language Sections and zero or more Data Element Name classes. The Language Sections document the data category for a specific working language. For each data category there shall always be an English Language Section with at least one Definition Section and one Name Section. The Data Element Name Sections document the names of a data category in one or more databases, formats or applications. The following attribute is associated with the Description Section:

— **+profile** [1..*]: attribute used to relate the current data category specification to one or several thematic domains treated by ISO/TC 37 (for example, morphosyntax, syntax, metadata, language description, etc.). When first created, the value of **+profile** for a data category specification defaults to -private-, unless or until the user selects one or more thematic domain profiles. Submission for standardization requires the selection of at least one thematic domain profile because it is the relevant TDG that is responsible for maintenance of standardized data category specifications.



△ inheritance relationship from subclass to superclass

♦ composition relationship between two classes

◊ aggregation relationship between two classes

Multiplicity values: 0..1 = 0 or 1; 0..* = 0 or more; 1 = 1 and only 1; 1..* = 1 or more instances may occur.

Figure 5 — Description part of the DCR data model

7.5.2 Information to be expressed in the Language Section

The Language Section describes the data category concept within the context of a given working language.

- **+language** [1]: the working language. The content of **+language** shall conform to IETF BCP 47, RFC 5646[17].
- **+note** [0..*]: any additional information associated with the data category, excluding technical information that would normally be described in **+explanation**.
- **+name section** [0..*]: records a possible name for the data category in a specific language. The Name Section can be repeated within a Language Section. The following descriptive attibutes are associated with the Name Section:
  - **+name** [1]: one-word or multi-word unit used to refer to the data category in the corresponding working language. Names given to a data category shall not be used for the purpose of identifying a data category (see **+identifier**);
  - **+name status** [1]: indication of name acceptability and usage, having the following value domain:
    - -standardized name-: the name that has been approved by a national, regional, or international standards body,
    - -preferred name-: in cases where multiple names exist, the name that has been specified as the most appropriate name by an authoritative body or for a specific environment or application,
    - -admitted name-: in cases where multiple names exist, the name that has been specified as acceptable by an authoritative body or for a specific environment or application,
    - -deprecated name-: a name that has been rejected for use by an authoritative body or for a specific environment or application,
    - -superseded name-: a name which the DCRB has confirmed to be no longer recommended for use in the DCR community and for which a successor name has been designated as preferred for use.
- **+definition section** [0..*]: definition of the data category concept associated with the data category, written in the language of the language section. The definition shall be provided for the mandatory English Language Sections of all data category specifications. Definitions are specified in the Definition Section, which contains the following attributes:
  - **+definition** [1]: definitive formulation that should be general enough to apply to all thematic domains and implementations of the data category;
  - **+source** [1]: source from which the definition has been borrowed or adapted;
  - **+note** [0..1]: any additional information about the definition.
- **+explanation section** [0..*]; further information about the data category concept. Explanations are specified in the Explanation Section, which contains the following attributes:
  - **+explanation** [1]: any additional information about the data category that would not be relevant for a definition (for example, more precise linguistic background concerning the use of the data category);
  - **+source** [0..1]: the source from which the explanation has been borrowed or adapted.
- **+example section** [0..*]: a sample instance reflecting the data category. These examples should be limited to those that illustrate the data category in general, excluding language specific usage, which should be documented in a Linguistic Section. Examples are specified in the Example Class containing the following attributes:
  - **+example** [1]: an example that illustrates the data category in general, excluding language specific usage;
  - **+source** [0..1]: the source from which the example has been borrowed or adapted.

7.5.3 Information to be expressed in the Data Element Name Section

The Data Element Name Section shall be used to record one name for the data category as used in a given database, format or application. Instances of the Data Element Name Section may be repeated within the Description Section in order to reflect different names that are used in different applications. The following attributes are associated with the Data Element Name Section:

— **+data element name** [1]: one identifier [word, multi-word unit or (alpha)numeric representation] used to refer to the data category in a given database, format or application. Data element names shall not be used for the purpose of identifying a data category outside a given database, format or application (see **+identifier**).

— **+source** [1]: information that identifies the database, format or application in which the data element name is used.

7.6 Conceptual Domain classes

7.6.1 Distinguishing conceptual domains

The DCR supports three kinds of conceptual domain: open, constrained and closed. An open conceptual domain imposes no restriction on the values that make up the conceptual domain as specified in the Open Conceptual Domain. A closed conceptual domain comprises an enumeration of permissible values which are selected from the DCR. A constrained conceptual domain specifies permissible values which cannot be expressed as a closed conceptual domain. An example of a constrained conceptual domain is: *all dates after 1965*. All these classes contain the following attribute:

— **+data type** [1]: the data type, as defined for W3C XML Schema[11] of this complex data category; the default data type is *string*.

7.6.2 Information to be expressed in the Open Conceptual Domain

An open conceptual domain allows all the possible values associated with a certain data type.

7.6.3 Information to be expressed in the Conceptual Domain Rule

Linguistic resources sometimes impose additional constraints using various schemas. The Conceptual Domain Rule allows users to express constraints on the possible values of a conceptual domain associated with a given data type in a rule language suitable for the schema in question. The same constraint may be expressed in several languages. The responsible TDG should confirm equivalence during the evaluation stage of the standardization process.

— **+rule type** [1]: the language used to express the rule, for example, W3C XML Schema or the Object Constraint Language[11][16];

— **+rule** [1..*]: the constraint expressed in the rule language.

7.6.4 Information to be expressed in the Value Domain

The Value Domain enumerates permissible values in the form of simple data categories.

— **+value** [1..*]: reference to a simple data category describing one element of the set of permissible values associated with a complex data category.

EXAMPLE The Value Domain for */grammaticalGender/* could be {*/masculine/*, */feminine/*, */neuter/*}.

7.6.5 Information to be expressed in the Profile Value Domain class

Complex data categories can be associated with multiple profiles. The Profile Value Domain subclass allows for the association of a certain value domain with a specific profile.

— **+profile** [1]: the profile with which this value domain is associated.

It should be noted that the constraint associated with the Closed Data Category as expressed in OMG's Object Constraint Language[16] imposes the following restrictions on the profile value domains:

a) there can only be one Profile Value Domain per profile;

b) for each profile of which a complex data category is a member, there should be a Profile Value Domain;

c) Profile Value Domains are only allowed for profiles of which the complex data category is a member;

d) only simple data categories that are associated with the same profile as a complex data category can be a members of its value domain;

e) Linguistic Sections can never extend the Value Domain of a Closed Data Category.

7.7 Linguistic Section classes

The Linguistic Section class is used to specify the behaviour of a complex data category in a specific object language. The superclass Linguistic Section provides the following information:

— **+language** [1]: the language being described (that is, the *object language*). The content of the **+language** attribute shall conform to IETF BCP 47, RFC 5646[17];

— **+conceptual domain** [0..*]: additional restrictions, for example additional constraints or a subset of the value domain, on the conceptual domain declared for a complex data category that are relevant to the object language described in the Linguistic Section;

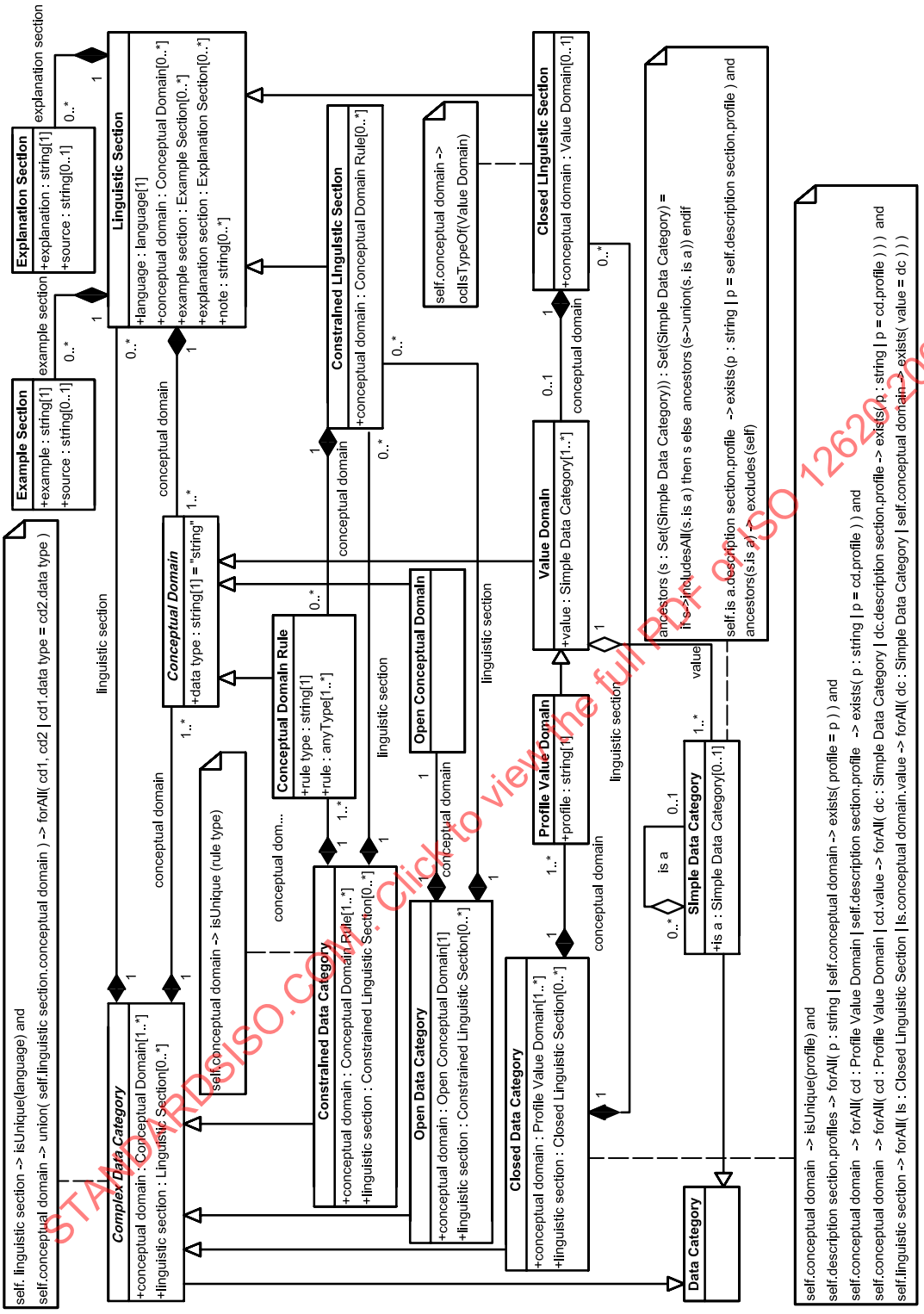
— **+example section** [0..*]: an example of how the data category is used for the current object language; the same Example Section class is used as described for the Language Section class;

— **+explanation section** [0..*]: an additional explanation specific to the use of the data category in the object language; the same Explanation Section class is used as described for the Language Section class;

— **+note** [0..*]: additional information, excluding technical information that would normally be described in **+explanation**.

Subclasses of the Linguistic Section class allow users to restrict the conceptual domain of a complex data category. The Constrained Linguistic Section allows the addition of object language-specific constraints to a constrained or open data category using one or more Schema-Specific Domains. These constraints should create subsets of the conceptual domains of the complex data categories, which is to say that they shall not be allowed to expand the set of permissible values.

The value domain of a closed data category may be restricted by using a Closed Linguistic Section. For instance, for the closed data category */grammaticalGender/*, the French Closed Linguistic Section could limit the value domain to {*/masculine/*, */feminine/*}. When the closed data category uses Profile Value Domains, the value domain of the data category for a specific combination of a profile and an object language is determined by the intersection of the value domains involved.



△ inheritance relationship from subclass to superclass

♦ composition relationship between two classes

◊ aggregation relationship between two classes

Multiplicity values: 0..1 = 0 or 1; 0..* = 0 or more; 1 = 1 and only 1; 1..* = 1 or more instances may occur.

Figure 6 — Linguistic part of the DCR data model

7.8 Referencing a data category

The explicit reference to a data category should 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, a data category such as */partOfSpeech/* may be referenced by a URI such as http://www.isocat.org/datcat/ISO-DC-1345.

Some schema languages have built-in constructs for embedding these PIDs. In the case of specifications expressed in the ODD language of the Text Encoding Initiative, the <equiv> construct should be used, see [10] and [18] for further reference. For instance, the following format will signal that the element being specified (<pos>) has the meaning defined for */partOfSpeech/* in the DCR:

```
<elementSpec ident="pos">
    <equiv name="partOfSpeech" uri="http://www.isocat.org/datcat/ISO-DC-1345"/>
    <!-- additional specifications here -->
</elementSpec>
```

Schema languages that lack these provisions, but which are still based on an XML vocabulary, can use a small DC Reference vocabulary to embed the data category PIDs. For instance, one could specify that a *POS* element is equivalent to */partOfSpeech/* in the DCR by embedding the dcr:datcat attribute at the appropriate location in the Relax NG Schema:

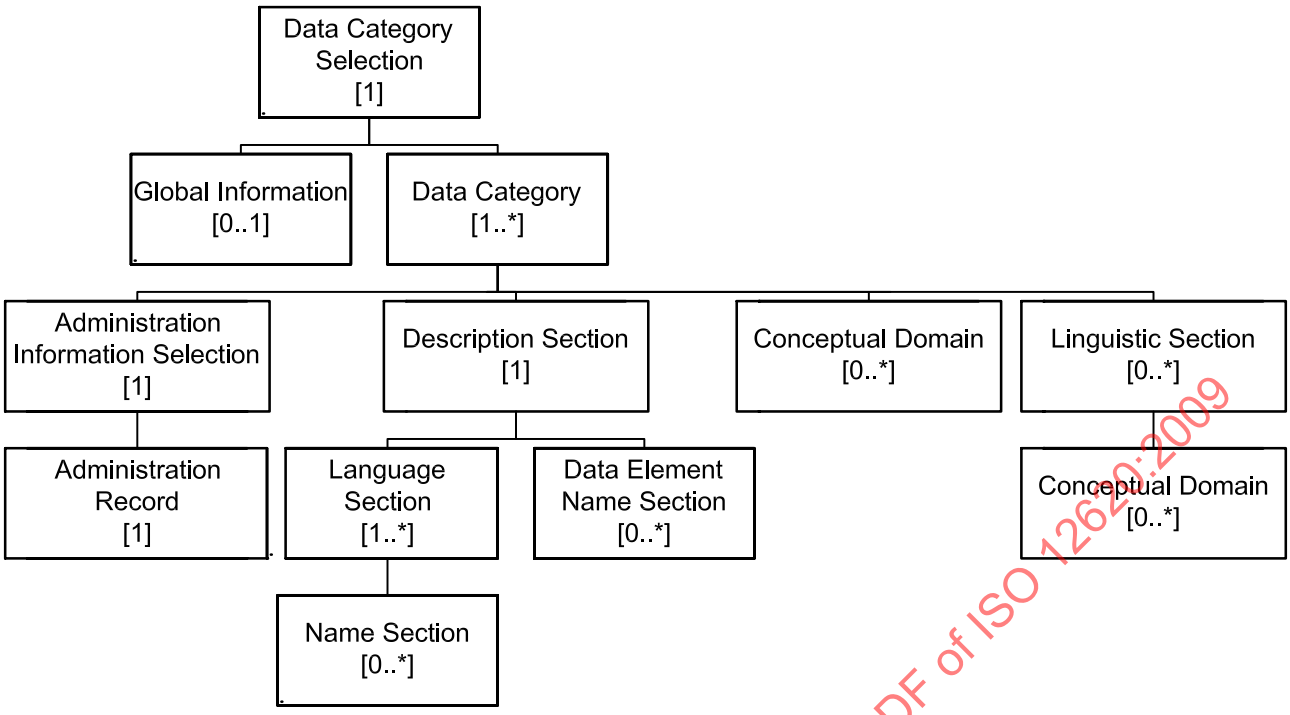
```
<rng:element name="POS" dcr:datcat="http://www.isocat.org/datcat/ISO-DC-1345">
    <!-- additional specifications here -->
</rng:element>
```

See Annex A and http://www.isocat.org for the DC Reference Relax NG schema for use in these cases.

7.9 Data Category Interchange Format

This subclause describes the Data Category Interchange Format (DCIF) used for archiving and exchanging all or part of the DCR within ISO/TC 37, as well as for applications where individuals have to manipulate and transmit their own proprietary data categories in the field of language resources.

The DCIF model shown in Figure 7 is strongly related to the class diagram that represents the DCR data model in Figures 3 to 6. However, as the class diagram is quite complex, the interchange format provides a simplified version of that model. The DCIF is described as a hierarchical component model. The components relate to the major classes of the data model. Simplification is achieved by subsuming information from some of the associated classes into the major classes. Where needed, the resulting components are annotated with a type attribute which indicates the original subclass in order to ensure lossless or near lossless round-trip conversions for DCIF documents. However, the DCIF schema allows a superset of the documents actually allowed by the DCR data model, for instance, the DCIF schema allows complex data categories to appear in a value domain. A DCIF export of a DCS from the DCR will always result in a DCIF XML document that complies with both the DCR data model and the DCIF schema. Due to the looser DCIF schema, the DCR should always validate a DCIF XML document against the DCR data model during import.



Multiplicity values: 0..1 = 0 or 1; 0..* = 0 or more; 1 = 1 and only 1; 1..* = 1 or more instances may occur.

Figure 7 — Data model underlying the DCIF

In some instances (data archiving, migration, etc.), the DCS expressed in a DCIF representation will comprise all the data category specifications in the DCR. In most cases, however, any given DCIF output will probably reflect a DCS for a TDG or for an individual expert or group of experts, any of which may comprise a DCS for an application if so desired.

In addition to the structural simplification of the DCR data model, the DCIF also includes the following simplifications:

— of the change histories saved in the Description Section, the DCIF only includes the creation and the most recent change descriptions;

— administration information internal to the DCR is excluded from the DCIF, which includes Registration Group, Submission Group, Stewardship Group and Decision Group, as well as **+administration status** and **+administration note** from the Description Section.

An example DCIF document is shown in Annex B. DCIF documents shall conform to the compact schema version shown in Annex C, as per ISO 19757-2:2003. The full DCIF schema is also available in electronic form at http://www.isocat.org/12620/.

8 Management procedures for the ISO/TC 37 DCR

8.1 General organization

Although the DCR comprises a central resource, it can be subsetted into groups of data categories, each associated with a thematic domain. Each of these subsets comprises a Data Category Selection (DCS) that is administered by a thematic domain group (TDG) made up of appointed experts. These DCSs also provide a logical framework for submitting and maintaining new data categories. Each individual data category specification contains an attribute indicating its profile value, which states the thematic domain or domains with which it is associated. In this regard, the management of the registry is not fully centralized, but is instead

based on a structure that draws on the relevant expertise distributed throughout the subfields of linguistic resources as represented in ISO/TC 37. Figure 8 illustrates the distribution of thematic domain groups within the administrative organization of the DCR.

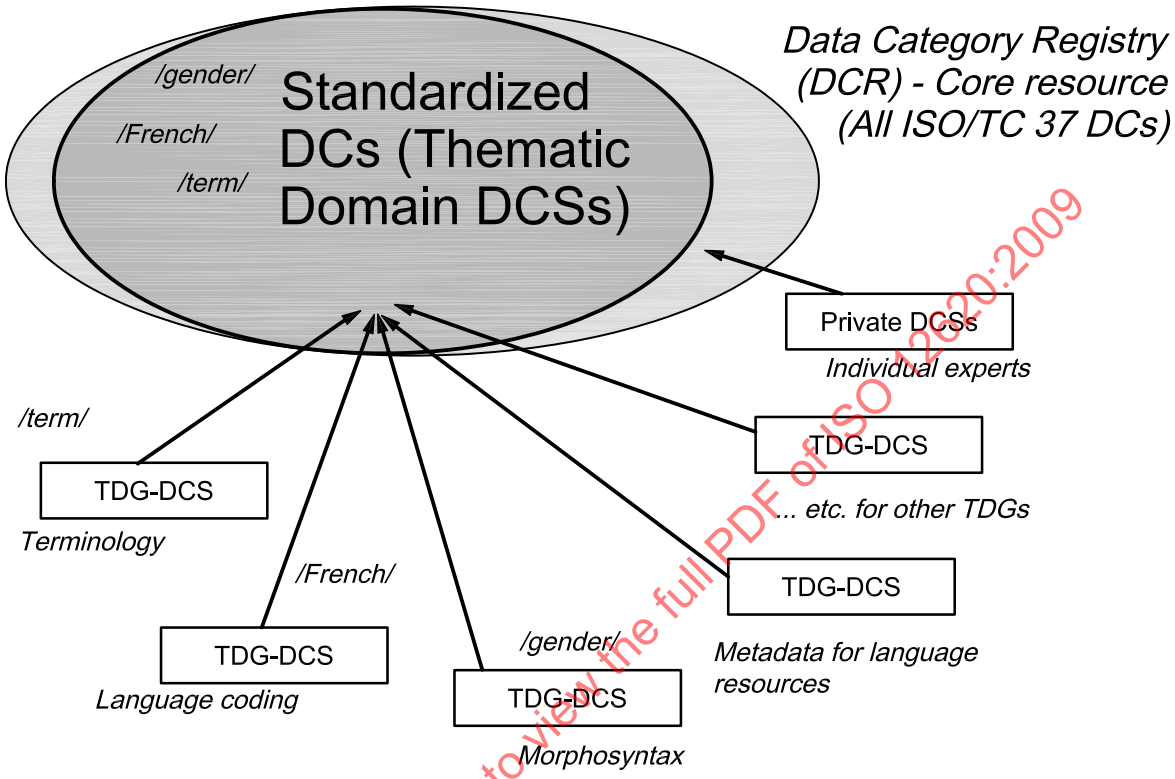


Figure 8 — Thematic domain-related Data Category Selections and private data categories maintained by individual experts

Although this illustration appears to imply that a data category such as */masculine/*, which is a simple data category and a part of the value domain for the closed data category */grammaticalGender/*, is "owned" by the Morphosyntax TDG, it obviously is used in the DCSs for a variety of TDGs. The outer oval in the drawing represents *Private* domain data categories, that is, those that have been created in the DCR by individual experts, but have not been submitted for inclusion in the standardized portion of the collection (see 8.4.3).

8.2 Roles and responsibilities

The DCR shall be freely available online at http://www.isocat.org for public consultation in order to ensure that language practitioners and system implementers will be able to use it systematically in any situation. Persons wishing to participate in the work of the DCR should declare their interest by filling in an online form, whereupon they are designated as *experts*. Independent experts are assigned their own private workspace within the registry where they may select existing data categories for inclusion in their own DCSs and create new data category specifications for their own private use. They can also submit data categories to be considered for inclusion in the global set of data category specifications as outlined below or propose modifications to existing data category specifications.

8.3 Thematic domain groups

Apart from the work of individual experts described in 8.2, different TDGs within ISO/TC 37 are responsible for submitting and selecting data category specifications designed to meet data documentation needs in specific

thematic domains. Viewed as a set, these data categories make up a TDG-specific DCS. In order to define such a DCS, a TDG shall be constituted under the following conditions:

a) ongoing standards development, or a new descriptive thematic domain that appears to be essential to the SC or TC in question can give rise to the need to define an additional DCS for a new thematic domain;

b) three months prior to the SC or TC plenary that will approve any new proposal(s), the SC or TC shall provide a document ("New Thematic Domain Group Proposal") that states the purpose and scope of the new Thematic Domain Group and its possible relations to existing TDGs associated with the DCR;

c) in response to this need, the SC or TC may, if it deems fit, pass a resolution establishing the TDG.

The TDG shall be composed of the following individuals, who will serve as judges and members of the Maintenance Team charged with the evaluation of data category specifications submitted to the TDG:

— a chair, appointed originally by the SC at the time the group is established, with subsequent appointments as necessary;

— a group of experts designated by the P members of the SC;

— a group of relevant outside experts proposed by the members or by the chair of the TDG if the expertise of the group needs to be augmented for the desired progress of its work. The total number of experts designated from outside the SC should not exceed 50% of the total number of experts in the TDG.

Once established, the TDG is also assigned a profile value, which is then created in the registry, reflecting the DCS required for that thematic domain. The value or values for the *profile* attribute is/are recorded in the relevant data category specification(s) and this information is used to generate the DCS as a list or set of data category specifications.

In some specific cases, for instance when a TDG is appointed in conjunction with a standard under development, there can be some variations to the above-mentioned procedures. For instance, it could be possible that an existing ISO 639 Registration Authority, together with its existing procedures and structure, could be designated as the TDG for language description with regard to the elaboration of a new part of ISO 639. In other cases, standardization activities will be conducted in close collaboration with other ISO committees, such as ISO/TC 46/SC 4 for language codes/descriptions; or ISO/JTC 1/SC 36 for CALL (Computer Assisted Language Learning) applications.

8.4 Working procedure

8.4.1 Data category decision process

The process that leads to the introduction or revision of a data category in the standardized portion of the DCR shall occur in four steps as outlined in 7.4.2, point 3 (+administration status), whereby the standardization process begins with the submission of a DC to a TDG:

0. a *creation process*, which involves the creation of a data category by an individual expert and is not itself a part of the standardization process because data categories remain in the expert's private workspace until and unless they are eventually submitted for standardization;

1. a *submission process,* which initiates the standardization process when the data category is submitted as a Change Request (CR) to a TDG for inclusion in the standardized portion of the DCR;

2. a *selection process*, during which the TDG identifies those data categories that are relevant for a certain application field within ISO/TC 37; this selection process parallels the *evaluation* stage in Annex ST as represented in Figure 9;

3. a *harmonization process*, overseen by the DCRB, which guarantees the coherence of new proposals within the scope of the DCR and data categories it already contains; this process, along with final approval of new data categories by the DCRB, parallels the *validation* stage in Annex ST as represented in Figure 9.

8.4.2 Creation of a new data category

Experts, acting either individually or on behalf of a TDG, create data category specifications in their own private workspaces and have the option of retaining them in their private DCS or submitting them for inclusion in the public, standardized portion of the DCR. As noted above, all newly created data categories are automatically assigned to the *Private* domain and remain outside the standards process unless explicitly submitted to a TDG.

8.4.3 Submission of a Change Request (CR)

8.4.3.1 Submission of a CR for a new data category

The submission to a TDG of a CR for a new data category shall be based on a description compliant with the DCR data model, with the following restrictions:

— information associated with the Description Section and its subordinate classes and attributes, which shall be documented during the *submission process,* including:

  — at least one definition in the English Language Section,

  — sources, if the definitions are taken from a known source,

  — at least one thematic domain profile value associated with the data category specification,

  — at least one English Name Section for the data category;

— a short description justifying the relevance of the data category to the field of language resources;

— mandatory administration information generated automatically by the DCR software;

— optional information, such as additional names and definitions in other languages, if applicable and available.

8.4.3.2 Submission of a CR for a modification proposal for an existing data category specification

Any expert or TDG can propose a modification to a data category. Such a modification request shall also be treated as a CR and include the following information:

— the data category uniquely identified with its persistent identifier;

— only the attributes for which a change is suggested (either a modification or a new piece of information);

— a short description justifying the proposed change.

8.4.3.3 Submission of a CR to assign an existing data category to a thematic domain DCS

Any expert or TDG can propose that an individual data category be added to the DCS of an additional thematic domain by adding the appropriate profile value in the data category specification. Such a request shall be treated as a CR and shall include the following information:

— the data category uniquely identified with its persistent identifier;

— a short description justifying the relevance of assigning the data category to the thematic domain;

— the name of the thematic domain to which the data category is to be added.

8.4.3.4 Submission of a CR for the selection of data categories making up a thematic domain DCS

The full set of data categories making up the DCS associated with a given thematic domain is determined by a combination of the procedures outlined above for creating new data categories, modifiying existing data categories and assigning existing data categories.

8.4.3.5 Submission of a CR in conjunction with harmonization activities

Harmonization of data category definitions and usage among the various TDGs shall be an ongoing goal of these procedures and shall be overseen by the DCRB. Harmonization requests are treated as submissions for a CR as part of the standard workflow procedure as illustrated in Figure 9. If a need for harmonization arises with respect to existing data category definitions, harmonization would take place in the context of a modification CR.

8.5 Data Category Registry Board (DCRB)

8.5.1 Constitution

The DCRB has the duty to ensure that the scope and the coherence of the registry are maintained. The Board shall play a harmonizing role with regards to the proposals that are submitted by experts and by the thematic domain groups.

The DCRB shall be composed as follows:

— a group of experts designated by the P-members of ISO/TC 37;

— a chair, appointed by the ISO/TC 37 plenary for a period of two years, which may be renewed once.

8.5.2 Working procedure

The DCRB, in conjunction with the RA, is responsible for validation of submissions and for publication of the standardized component of the DCR.

8.5.2.1 Validation of a data category proposal

Any submitted data category shall be defined by at least the minimum necessary data category criteria as outlined in 8.4.3.

The data category shall be evaluated by the responsible TDG as reflected in the **+administration status** attribute outlined above in 7.4.2, approved, rejected, or referred to a different TDG.

The data category shall be validated at the level of the DCRB.

Workflow schedules in compliance with Annex ST of the ISO Supplement to the ISO/IEC Directives [9] are posted on the ISOcat website.

If the validation ballot receives a positive vote of more than 70%, the status of the data category shall be upgraded to “standardized”. If less than 70% is received, the data category shall be given a “rejected” status and reasons for rejection shall be directed to the submitter. If a rejected data category is re-submitted following modification, this re-submission shall follow the process for a new data category specification, but shall contain notes regarding the previous submission.

8.5.2.2 Publication of a reference version of the registry

Every six months, the Registration Authority charged with maintaining the DCR shall issue an updated version of the DCR including all standardized data category specifications that have been approved by the DCRB. This data snapshot version shall be assigned a date and a version number and shall be considered to be the official standard reference for the following six-month period. The DCRB shall inform ISO/TC 37 Member bodies and liaisons as well as the SC secretary and TDG chairs of the changes that have occurred in comparison to the previous issue (additions, modifications, and deprecations). Thus, those data category specifications that are “published” in this way shall comprise a “standard as database” of approved data categories, which nevertheless shall also remain available to all users in the dynamic environment of the DCR.

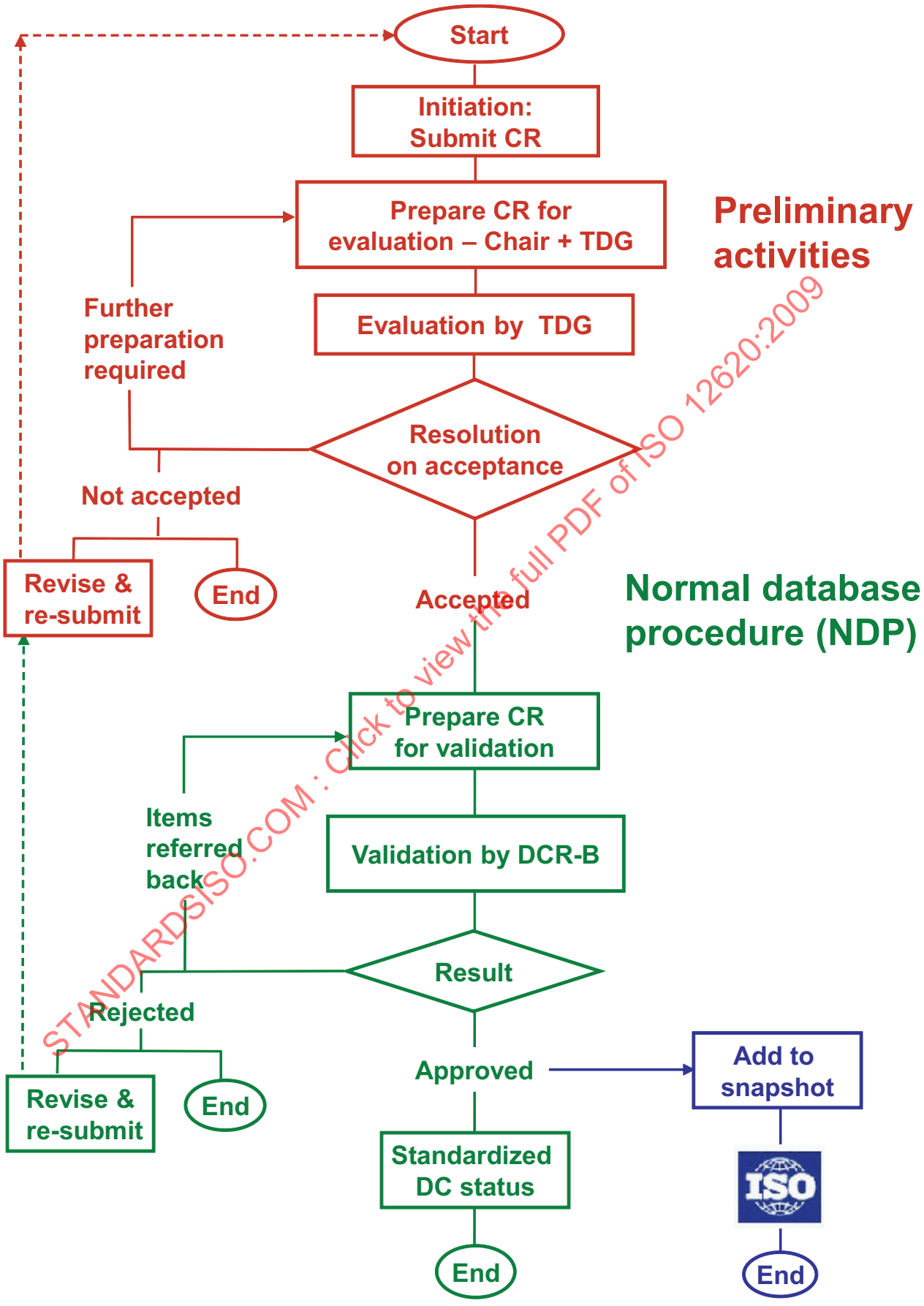


Figure 9 — Workflow diagram

Annex A
(normative)

Compact DC Reference RELAX NG Schema

```
default namespace dcr = "http://www.isocat.org/ns/dcr"

start = any

any = dcr_elements | element * - dcr:* { content }

content = ( dcr_attributes | attribute * - dcr:* { text } | any | text )*

dcr_attributes = dcr_attribute_datcat | dcr_attribute_any

dcr_attribute_datcat = attribute datcat { xsd:anyURI }

dcr_attribute_any = attribute dcr:* - dcr:datcat { text }

dcr_elements = dcr_element_datcat | dcr_element_any

dcr_element_datcat = element datcat { attribute pid { xsd:anyURI } }

dcr_element_any = element dcr:* { content }
```

NOTE This schema is open in the http://www.isocat.org/ns/dcr namespace; see the dcr_attribute_any and dcr_element_any named patterns. New attributes and elements introduced using this feature are intended to be used for DCR-specific information only.

Annex B
(informative)

Example of a DCIF Representation

The following example is a DCIF XML representation of the core information associated with a data category on the basis of the principles described in this document:

```
<?xml version="1.0"?>
<dataCategorySelection xmlns="http://www.isocat.org/ns/dcif" dcif-version="1.0">
   <globalInformation>
      Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
   </globalInformation>
   <dataCategory pid="http://www.isocat.org/datcat/ISO-DC-1234" type="complex">
      <administrationInformationSection>
         <administrationRecord>
            <identifier>grammaticalGender</identifier>
            <version>1.0.0</version>
            <registrationStatus>standard</registrationStatus>
            <origin>ISO 12620:1999</origin>
            <justification>
               In many, but not all, Indo-European languages grammatical gender
               describes the classification of nouns according to their behaviour
               with respect to inflection, pronoun agreement, semantics, morphology,
               and linguistic convention.
            </justification>
            <creation>
               <creationDate>1999-01-01</creationDate>
                  <changeDescription xml:lang="en">
                     Initial creation of the /grammatical gender/ data
                     Category
                  </changeDescription>
            </creation>
         </administrationRecord>
      </administrationInformationSection>
      <descriptionSection>
         <profile>terminology</profile>
         <languageSection>
            <language>en</language>
            <definitionSection>
               <definition xml:lang="en">
                  A grammatical category that indicates grammatical relationships
                  between words in sentences.
               </definition>
               <source>ISO 12620:1999</source>
               <note>
                  The concept of gender varies from language to language and is not a
                  universal feature of all languages.
               </note>
            </definitionSection>
            <exampleSection>
               <example xml:lang="en">
                  In French, vie (life) is feminine and is used with feminine articles
                  such as la, the feminine pronoun elle, and feminine adjective
                  endings, for example, une vie longue.
               </example>
            </exampleSection>
            <nameSection>
               <name>grammatical gender</name>
               <nameStatus>standardized name</nameStatus>
            </nameSection>
         </languageSection>
```

```
      </descriptionSection>
      <conceptualDomain type="closed">
         <dataType>string</dataType>
         <profile>terminology</profile>
         <value pid="http://www.isocat.org/datcat/ISO-DC-1111">
            <!-- PID reference to the /masculine/ simple DC -->
         </value>
         <value pid="http://www.isocat.org/datcat/ISO-DC-1112">
            <!-- PID reference to the /feminine/ simple DC -->
         </value>
         <value pid="http://www.isocat.org/datcat/ISO-DC-1113">
            <!-- PID reference to the /neuter/ simple DC -->
         </value>
         <value pid="http://www.isocat.org/datcat/ISO-DC-1114">
            <!-- PID reference to the /other/ simple DC -->
         </value>
      </conceptualDomain>
      <linguisticSection type="closed">
         <language>fr</language>
         <conceptualDomain type="closed">
            <dataType>string</dataType>
            <value pid="http://www.isocat.org/datcat/ISO-DC-1111">
               <!-- PID reference to the /masculine/ simple DC -->
            </value>
            <value pid="http://www.isocat.org/datcat/ISO-DC-1112">
               <!-- PID reference to the /feminine/ simple DC -->
            </value>
         </conceptualDomain>
      </linguisticSection>
   </dataCategory>
</dataCategorySelection>
```