ISO TC 37/SC 3 N 488

Date: 2003-10-31

ISO/CD 12620-1

ISO TC 37/SC 3/WG 4

Secretariat: DIN

Terminology and other language resources — Data categories — Part 1: Specification of data categories and management of a data category registry for language resources

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: International Standard
Document subtype:
Document stage: (30) Committee
Document language: E

M:\dms\dtec\xdom741\iso\12620\Part 1\N488 ISO CD_12620-1_(E).doc STD Version 2.1c

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 12620-1 was prepared by Technical Committee ISO/TC 37, *Terminology and other language resources*, Subcommittee SC 3, *Computer applications in terminology*.

This second edition cancels and replaces the first edition (ISO 12620:1999), [clause(s) / subclause(s) / table(s) / figure(s) / annex(es)] of which [has / have] been technically revised.

ISO 12620 consists of the following parts, under the general title *Terminology and other language resources — Data categories*:

— *Part 1: Specification of data categories and management of a data category registry for language resources*

— *Part 2:Terminological data categories*

— *Part 3: ....*

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 standard as *data categories*, a designation commonly used in the TC 37 environment that reflects a type vs. token relation. Data categories as cited in TC 37 standards correspond to data element concepts in the ISO/IEC 11179 series of standards. Differences in approach among different 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 within the same resource domain (e.g., among terminological resources, lexicographical resources, annotated textual 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 resource domain have also be uniform in order to ensure interoperability of individual data category registries.

The creation of a single global data category registry for all types of language resources treated within the TC37 environment provides a unified view on the various applications of such a reference resource. Such a universal registry would include traditional collections, such as the current ISO 12620 or its successor standards, but could incorporate a wide range of other current projects, including perhaps as yet unforeseen data collections. The following applications may be relevant to consider as definable subsets of a language resource data category registry:

- Terminological data collection — the ISO 16642 standard explicitly refers to ISO 12620, which describes a set of reference data categories for terminology representation. Some of the data categories already defined in ISO 12620, include general-purpose management data categories (e.g., */source/, /responsibility/, /date/*, etc.) as well as linguistically oriented ones (e.g., /part of speech/). These data categories are relevant to a variety of different language resources, not just to terminology management. In the context of its current revision, such a collection could be incorporated at an early stage of the implementation of a single language resource registry;

- On-going and future activities within TC37/SC4 — a data category registry is intended to be the basis for the work items planned in the SC4 Business Plan. For instance, it will serve as a reference for the descriptors that would be used at various levels of linguistic annotation (morpho-syntactic, syntactic, discourse level etc.), for lexical representations (NLP lexica, Machine translation dictionaries, etc.), or for specific applications such as metadata for language resources, query languages or multilingual data representation (translation memories);

- Language codes — ISO 639-1 and ISO 639-2 contain codes for about 650 languages. The current work being done in TC37/SC2 will extend this number by an order of magnitude, with a clearer separating between the description of the language and its coding proper. Besides, having a reference set of language identifiers is an essential element of any linguistic annotation or representation scheme. It is thus natural that a future language resource data category registry be the background for the evolution of ISO 639.

- Lexicographical data — The deployment of a data category registry could accompany on-going activities within TC37/SC2 on the description of lexicographic data. This would ensure in particular that the formats used for describing lexicographical (SC2), terminological (SC3) and NLP oriented (SC4) data are comparable;

The Data Category Registry would eventually contain all data categories, with their complete history, data category description, and attendant metadata. Individual parts of the 12620 standard could then

specify the Data Category Selection (DCS) required for documenting linguistic resources within a specified thematic domain (e.g., terminology, lexicography, etc.).

It may not be seen as a priority to define such an ontology within TC37/SC4 in a short term perspective, and it may be wiser to wait for a stronger community to crystallize at an international level. Still, no choice should be made in the definition of the data category registry that would hamper further work in this direction.

This document is intended to provide a background on the various issues that have to be considered in order to implement a global data category registry in the context of ISO technical committee 37 (Terminology and other language resources) that can be used for the full range of language resources. More precisely, this document addresses the following issues:

- The role of data categories, not only in the domain of terminologies, but also for use with other language resources;

- The possible requirements that can be identified from the points of view of information content and overall management;

- A description of the possible organization of the data category registry;

- An overview of the possibilities that could lead to the proposal of an interchange format for data categories positioned at a cross-application level;

- A proposal of a possible interchange format for data categories, DCIF (Data Category Interchange Format), defined according to a methodology consistent with ISO 16642, which specifies an linguistic format as the combination of a metamodel and a selection of data categories.

Terminology and other language resources — Data categories — Part 1: Specification of data categories and management of a data category registry for language resources

1 Scope

This International Standard gives guidelines on the constraints related to the implementation of a data category registry applicable to all types of language resources, e.g., terminological, lexicographical, corpus-based, machine translation, etc. It specifies mechanisms for selecting and maintaining categories and specifies 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 8879:1986, (SGML) as extended by TC2 (ISO/IEC JTC 1/SC 34 N 029:1998-12-06) to allow for XML.

ISO/IEC 11179-3:2003, Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes.

ISO 16642:2003, Computer applications in terminology – TMF (Terminological Markup Framework).

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 11179 and the following apply.

3.1
conceptual domain
a set of valid value meanings
[ISO/IEC 11179-3]
3.2
data category
result of the specification of a given data field
[ISO 1087-2:2000, definition 6.14]
note: a data category is to be used as an elementary descriptor in a linguistic structure or an annotation scheme
note: a data category corresponds to a data element concept in ISO/IEC 11179
example: /part of speech/, /grammatical gender/, /grammatical number/, /feminine/, /plural/, /ablative case/
3.3
DE
data element
a unit of data for which the definition, identification, representation and Permissible Values are specified by means of a set of attributes
[ISO/IEC 11179, definition XXX]

3.4
DEC
data element concept
a concept that can be represented in the form of a Data Element, described independently of any particular representation
[ISO/IEC 11179, definition XXX]
3.5
DCS
data category selection
component of a TML's specification that constrains its informational content
[ISO 16642, definition 3.4]
3.6
data category instance
implementation of a data category in a specific application or coding system
note: ISO 639-1 represent a typical case of a set of data category instances corresponding to a DCS of language representations
3.7
data category specification
[ISO 12620]
3.8
DCR
data category registry
data category specification used as a normative reference for the description of a TML
[ISO 16642, definition 3.3]
3.9
object language
language being described
[ISO 16642, definition 3.10]
3.10
thematic domain
class of applications identified by the similarity of the data structures they need to manipulate
3.11
thematic domain committee
committee of expert in charge of selecting the data categories that are relevant for a thematic domain
3.12
thematic profile
thematic domain to which a data category is attached
NOTE a data category may have several thematic profiles
3.13
stewardship (of metadata)
the responsibility for the maintenance of Administration Records applicable to one or more Administered Items
NOTE The responsibility for the registration of metadata may be different from the responsibility for stewardship of metadata.
[ISO/IEC 11179-3, definition xxxx]
3.14
working language
language used to describe objects
[ISO 16642, definition 3.21]

4 Role of data categories in language resource management

4.1 Overview

Data category specifications are needed to identify the individual information units making up a data collection, or annotation scheme, for a given language resource. Figure 1 below shows possible uses for a set data category specification, i.e. a data category selection (henceforth DCS). As exemplified in ISO 16642, *Terminological Markup Framework (TMF)*, a data category selection is needed in order to define, in

combination with a meta-dodel, the various constraints that apply to a given domain-specific information structure or interchange format (e.g. expressed in XML). These constraints can be typically expressed as a DTD, a RelaxNG schema or an XML schema that will allow a computer application to check the validity of a language resource data collection against the intended specifications or to utilize a set of XSLT filters that will map the collection from one markup language to a neutralized dumped format (such as GMT in the case of ISO 16642) for archiving purposes and back, with the global purpose of mapping one markup language or format to another.

From a wider perspective, a formal model for representing data categories must 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 data category registry, which can be published either as a paper document (such as the printed version of ISO 12620-2) or an electronic resource, such as the global Data Category Registry (DCR) for TC 37 language resources. Typically, the designers of a given markup language or data management system will query such a registry in order to create their individual application profiles by selecting a subset of data category specifications from the global DCR. As a consequence, the formal representation of a data category shall comprise the specific attributes that document it (e.g., the data category name, definition, examples, comments, etc.). It shall also provide the context for its creation and management within a given registry.

Finally, providing a precise description of the data categories used within a given data collection in reference to certified registries allows for a quick diagnosis of the compatibility of this collection with any particular computer application and thus acts as metadata for this collection.

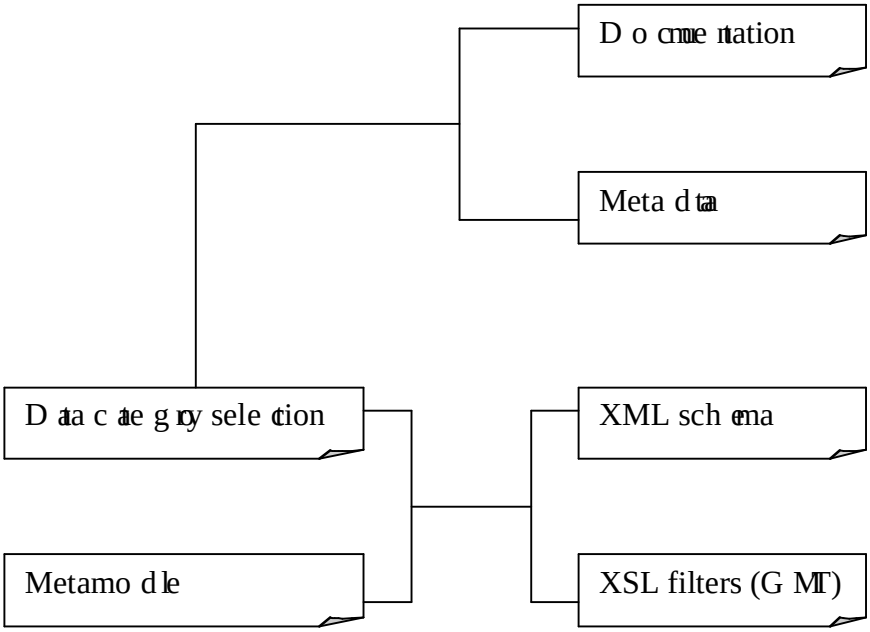

Figure 1 — The role of data category selections in the context of the definition of linguistic annotation schemes.

Figure 1 presents the notion of a *data category selection (DCS)*, e.g., the choice of a specific set of data categories taken from the global data category registry for use in a specific thematic domain within the framework of language resources. The diagram exemplifies the various roles of a data category selection in the process of defining and using any linguistic annotation scheme. A DCS is primarily intended to contribute to the specification of an annotation scheme in combination with a metamodel that expresses the general organization of a data model. As stated in ISO 16642 (TMF), such a selection guarantees a certain degree of interoperability between two data structures by comparing the selected data categories as well as the constraints that bear on them, in particular the nodes of the metamodel where each category is allowed to occur in the two data structures.

If the specification also contains the provision of styles and vocabularies (cf. 16642) for each data category, the DCS then contributes to the definition of a full XML information model which can either be made explicit through a schema representation (e.g. a W3C XML schema), or by means of filters to and from the GMT representation.

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

Furthermore, the data category selection 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.

4.2 A variety of data category selections (DCS)

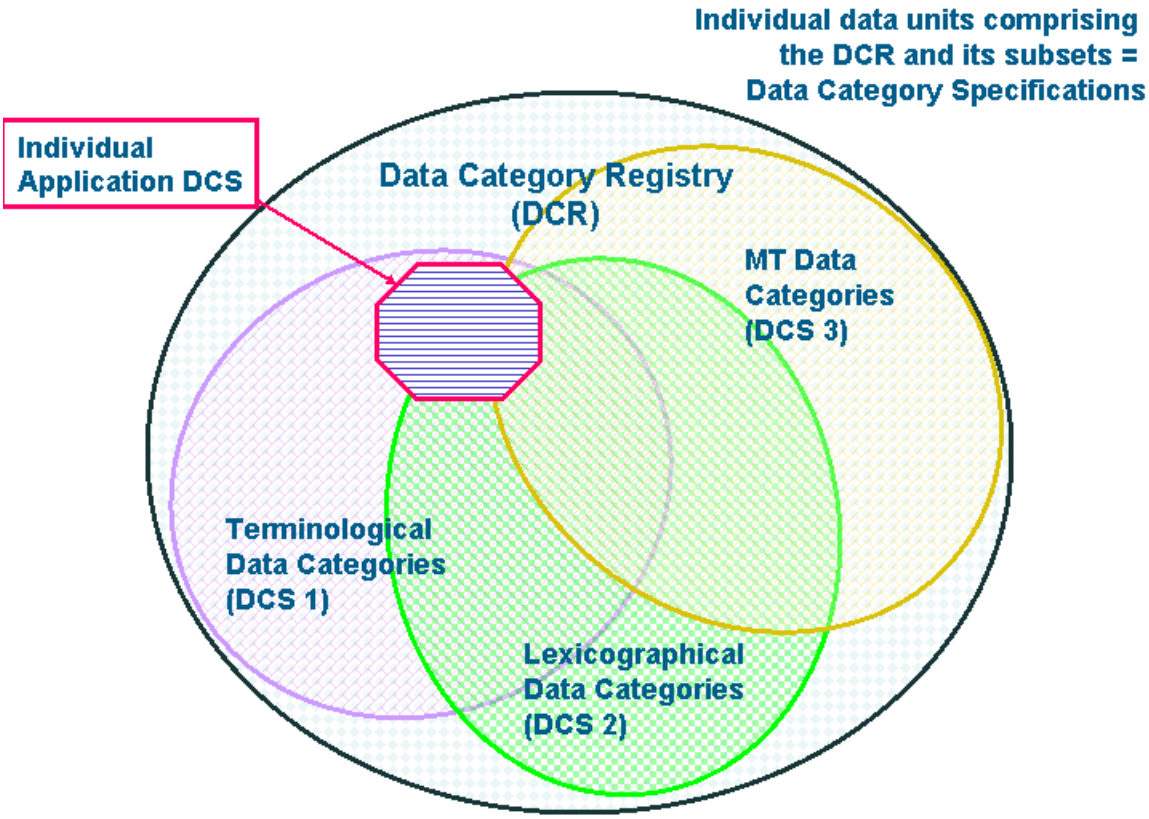

Figure 2 — The Data Category Registry (DCR) and its relationship to some possible Data Category Selections (DCS) associated with individual thematic domains (e.g., terminology, lexicography, machine translation, etc.)

Figure 2 above illustrates the relationship between data category specifications, the DCR, and any one of the possible DCSs that can be subsetted from the DCR. The patterned cells represented in the drawing correspond to individual data category specifications, each describing a given data category concept using the data category attributes set down in ISO 12620-1 with reference to the attributes defined in ISO/IEC 11179. Some data categories included in the DCR for *terminology and other language resources* are pertinent to a single thematic domain within this field. For instance, a *concept identifier* is probably unique to terminological resources (although not prescriptively), or a *sense number* is probably specific to lexicographical resources. Nevertheless, many data categories, frequently those of a strictly linguistic nature such as *part of speech,*

grammatical gender, grammatical number, 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 token relationship in different kinds of resources. Hence each thematic domain contributes all its data categories in the form of data category specifications to the global Data Category Registry, while at the same time identifying those data categories that it shares with other kinds of resources. A standard listing the subset of data categories used in a thematic domain will comprise a domain-specific Data Category Selection (DCS) taken from the DCR. The oval shapes in the Venn diagram represent such DCS subsets. A further, smaller subset can be selected from the domain DCS for use in a given application or collaborative environment. The octagon represented in figure 1 represents such a smaller subset. Note that while some of the data categories contained in this subset are common to several different kinds of language resources, this particular application is wholly contained within the DCS for terminological entries, so we can conclude that it is designed for use with a terminological application.

5 Requirements applying to the implementation of a data category registry for the language resource domain

In this section, we try to outline the basic requirements that a data category registry should fulfill to fit the needs of the various activities related to the scope of standardization activities within TC37.

We consider that the data category registry for TC37 shall:

- Be a reference for all the existing or future standards in TC37 related to data modeling or data interchange. This encompasses current activities in SC2 (language description and coding, lexicography), SC3 (terminology description) and any on-going and future SC4 work item (e.g. POS annotation);

- Register existing practices by associating a data category with the way it is implemented in specific projects or initiatives (encodings). This may consist in registering various types of encodings, from basic codes ('f' for feminine in Eagles morpho-syntactic descriptions) to actual XML implementation;

- Provide names and reference definitions in a variety of languages;

- Describe the usage of a data category in a variety of language settings. This may consist of a specific definition (for instance when the data category has a slightly application scope), some usage notes, examples, or list of values (e.g. the conceptual domain of /gender/ is {/masculine/, /feminine/} in French, and {/masculine/, /feminine/, /neuter/} in German;

- Describe the usage of a data category in a variety of data processing environments; e.g., some data categories function somewhat differently in machine translation lexica from the way the function in terminological resources or in human-oriented lexicographical resources;

- Associate administrative information to each data category so that it is possible to trace the submission, acceptance or revision of the data category;

- Associate a data category with one or several profiles corresponding to the application domains where the category is relevant (for instance, /Part of speech/ is relevant for POS annotation and lexical representation;

- Provide a mechanism by which a working group in TC37 can submit a group of categories relevant to their scope of activities;

- Be updated on a regular basis by integrating, according to rules to be defined, proposals from experts in the field;

- Provide a personal working space within which experts can upload and publicize their data category proposals, even before they are submitted to the registry;

6 Current situation in data category representation

[As background information, this section is kept here as is for the time being but should obviously be dissolved soon partly in the introduction and partly in the next section.]

6.1 General background: ISO/IEC 11179

ISO/IEC 11179 (“Metadata registry”) is a standard that has been developed within ISO committee JTC1/SC32 to provide a background for the description of data-elements. ISO/IEC 11179, Part 3 (4.13.1.1) states that, “A Data Element is considered to be a basic unit of data of interest to an organization. It is a unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes.” Indeed, data elements can take various forms, from a database field to an XML object in a complex structure. For instance, the Dublin Core fields are described (see …) using the original set of attributes provided by ISO/IEC 11179-3 in its first version.

Analogous to the relationship between terms and concepts, data elements (DE) represent data element concepts (DEC). The DEC comprises an abstraction, the notional mental construct referenced by a data element name. Besides, ISO/IEC 11179 provides means to describe the values associated to a data element. To be more precise, a data element can be associated to a value domain, which reflex the association that is expressed at a conceptual level between a data element concept and a conceptual domain (see figure 3).

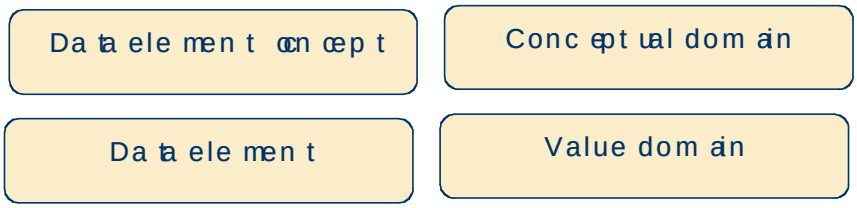

Figure 3 — ISO/IEC 11179-3 basic concepts

The prototypical application of this data organization in the field of language resources will be to describe, on the one hand, the abstract notion of a data category together with its possible values, as a Data Element Concept (linked to a Conceptual Domain), and, on the other hand, any possible instantiation as a Data Element (linked to a Value Domain). For instance, the data category registry may contain:

- at the DEC level: /Grammatical number/ (Conceptual Domain: {/Singular/, /Plural/})
- at DE level: XML attribute ‘num’, with its domain values consisting of {‘s’, ‘p’}

It should be noted here that for sake of homogeneous representation, /Singular/ and /Plural/ should also be representing as DECs, with ‘s’ and ‘p’ described as DE and respectively linked to them as possible instances. Figure 4 shows a similar case with the data category /Gender/.

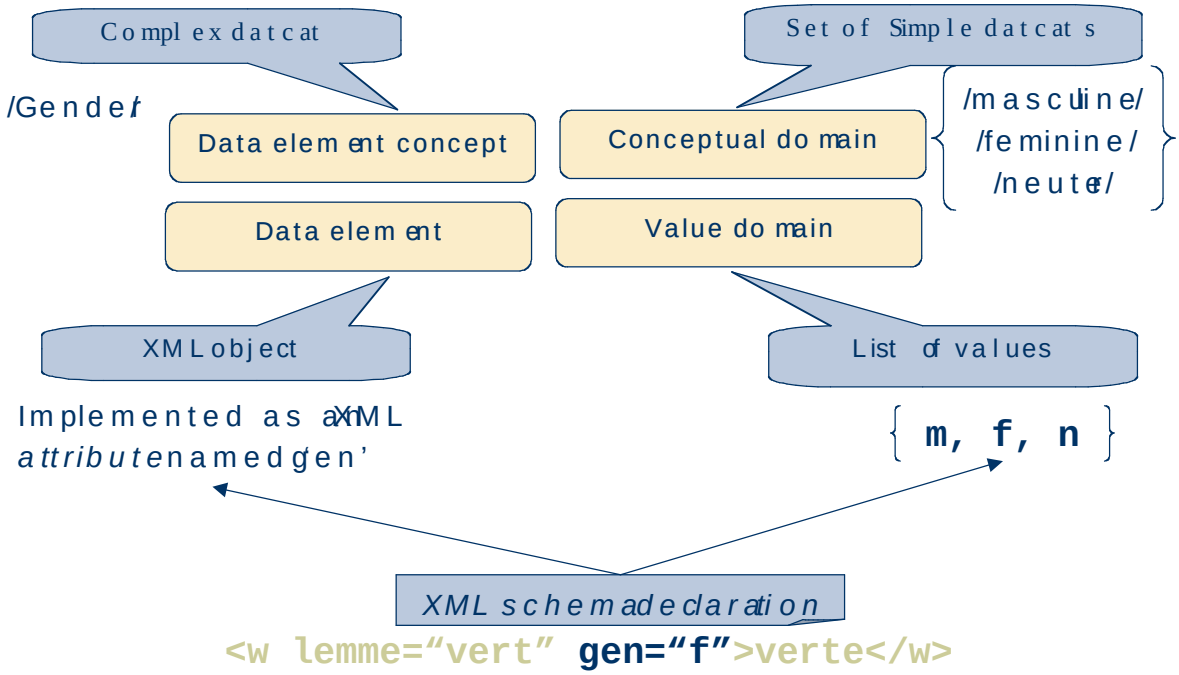

Figure 4 — Using ISO/IEC 11179 concepts for Data Category descriptions.

Another difference between a DEC and a DE is that we can dissociate the description of a concept from a possible assignment of a code to it. For instance, the two codes 'fr' and 'fra', taken from ISO 639-1 and ISO 639-2, have been defined to refer to exactly the same language, i.e., French. In ISO/IEC 11179 terms, we could define the DEC /French/, to which would be attached two DEs, corresponding to ISO 630-1 and ISO 639-2 implementations.

This property could also be used in the language resource registry to document existing implementations of data categories in various projects or consortia. For instance, data-categories used for the metadata description of language resources could be mapped onto the IMDI and OLAC vocabularies, thus making explicit the possible mappings between the two initiatives.

6.2 Main issues related to the new version of ISO/IEC 11179-3

The new version of ISO/IEC 11179-3 provides a sound background for the management of DECs and DEs altogether by introducing a general notion of *Administered Item*. Administered items as described are made of two parts:

- the Administration and Identification region supports the administrative aspects of Administered Items in a registry. This region addresses in particular the identification and registration of items submitted to the registry; the organizations that have submitted items to the registry, and/or that are responsible for items within the registry, including Registration Authorities; contact information for organizations; supporting documentation; relationships among administered items (see section 4.8 of 11179-3 (rev.));

- the Naming and Definition region which, following exchanges between TC37 experts and JTC1/SC32 experts is somewhat similar to a terminological entry;

We will come back later to the role of terminology principles for the representation of data category. We explore here the main elements that we would like to retain from ISO/IEC 11179-3 (rev.) in the domain of administration and identification of administered items. The entry point to this information is a single level: the Administration Identification (AdminIdent) level.

7 An interchange format for data categories in TC37

7.1 Introduction

This section describes DCIF, the Data Category Interchange Format, to be used as the underlying tool for archiving and exchange all or part of the data category registry within TC37, but also for applications where individuals have to manipulate and transmit their own proprietary data categories in the field of language resources.

DCIF is described as a general model, which can in turn be implemented using the GMT format, using the methodology presented in ISO 16642. The GMT DTD is made available in annex A.

7.2 General principles

The main principle that lies at the root of DCIF is to combine the view of a data category as an administered item as in ISO/IEC 11179-3 complemented by a descriptive component inspired by the terminological metamodel described in ISO 16642 to provide a sound basis for representing language dependant information.

Note: in the following sections, the data categories associated with the DCIF meta-model are marked with occurrence indicators (in square brackets) as follows:

- 'one and only one' states that the data category is mandatory and shall not be repeated;
- '?' states that the data category is optional and shall not be repeated;
- '+' states that the data category is mandatory and may be repeated;
- '*' states that the data category is optional and may be repeated.

7.3 Metamodel

The following metamodel corresponds to the fusion of the main structure of an administered item together with a three-level documentation part that partially maps a portion of the ISO 16642 metamodel.

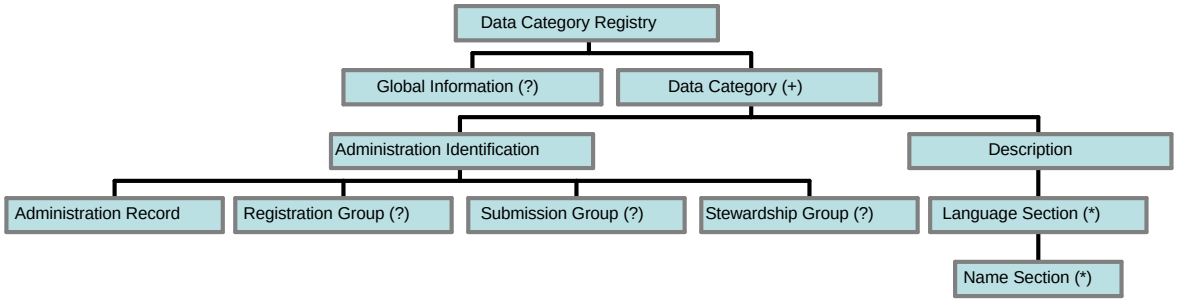

Figure 5 — The metamodel underlying DCIF (occurrence markers are the same as for data categories, except that no marker is equivalent to ' one and only one ').

A Data Category Registry shall consist of two main parts:

- A Global Information level providing means to specify the context of the interchange process[1];
- A list of at least one Data Category nodes.

[1] Question: should this level be left opened or should we go deeper in its description?

Each Data Category shall consist of two mandatory sub-levels:

- One dedicated to the administration and identification of the data category (Administration Identification level);
- One dedicated to the linguistic and conceptual documentation of the data category (Description level).

The Administration Identification level and the Description level are described in the following sections.

7.4 The Administration identification level

7.4.1 General

The administration identification (AI) level can be further decomposed into four sublevels:

- Administration Record (AdminRec): groups together the information associated with the global management of the administered item;
- Registration Group (RegGrp): contains the information related to the registration authority in charge of the administered item. For the registry being described here, this registration authority may globally be ISO committee TC37, or a specific sub-committee;
- Submission Group (SubGrp): contains the information related to the entity that has submitted the data category to the registry. This may either be the thematic committee that has selected the data category, and/or possibly the expert at the origin of the submission;
- Stewardship Group (StewGrp): contains the information identifying the entity responsible for the maintenance of the administered item (for instance, the body or institution in charge of the registry).

Note that if the TC 37 DCR is maintained by one single entity, the Stewardship Group shall always be identical from one data category to another.

7.4.2 Information to be expressed at administration record level

The administration record level is centered on the identification and maintenance of the administered item, it can be associated to the following fields:

- /identifier/ [one and only one; ISO/IEC 11179-3]: which uniquely identifies the data category in the registry, under the condition that it is refined by /registration authority/ and /version/. The combination of the three descriptors provides a unique key to the right version of the data category in its conditions of registration;

  Note: according to ISO/IEC 11179, 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 (e.g. /term/, /normative authorization/, /preferred term/), but such a practice should not preclude the usage of additional names for the data category in English or any other language.

- /registration authority/:
- /version/: used to refine /identifier/ to indicate the version of the data category;
- /administration note/ [?; ISO/IEC 11179-3]: any general note about the Administered Item;
- /administration status/ [one and only one]: “a designation of the status in the administrative process of a Registration Authority for handling registration requests. NOTE: The values and associated

meanings of ‘administrative status’ are determined by each Registration Authority. C.f. ‘registration status’.

- /registration status/ [one and only one; 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:

  - /standard/: the Registration Authority confirms that the administered item is of sufficient quality and of broad interest for use in the Registry community;
  - /qualified/: the Registration Authority has confirmed that the mandatory metadata attributes are complete and conform to applicable quality requirements;
  - /candidate/: It has been proposed for progression up the Registry registration levels;
  - /retired/: the Registration Authority has approved the administered item as no longer recommended for use in the registry community and should no longer be used;
  - /superseded/: the Registration Authority has approved the administered item as no longer recommended for use in the registry community but the successor administered item is the preference for uses.

- /creation date/ [one and only one]: the date when the data category has been initially created (for instance in an expert’s working space/private area);
- /effective date/: [?] “the date an administered item became/becomes available to registry users” (ISO/IEC 11179-3)
- /last change date/ [?]: the date when the data category has last undergone a change (see change description);
- /change description/ [? (mandatory if /Last change date/ is used)]: free text description of the modification undergone by the data category (e.g. “definition updated…”)[2];
- /explanatory comment/ [*; ISO/IEC 11179-3]: descriptive comments about the Administered Item;
- /origin/ [?; ISO/IEC 11179-3]: source (document, project, discipline or model) for the Administered Item;
- /unresolved issue/ [*; ISO/IEC 11179-3]: problem that remains unresolved regarding proper documentation of the Administered Item;
- /until date/ [?; ISO/IEC 11179-3]:the date an Administered Item is no longer effective in the registry;

7.4.3 Information to be expressed at registration group level

- /organization Name/ [one and only one]

7.4.4 Information to be expressed at submission group level

- /organization Name/ [one and only one]
- /contact/ [*]

---

[2] The data category registry should keep track of all changes.

7.4.5 Information to be expressed at stewardship group level

- /organization Name/ [one and only one]
- /contact/ [*]

7.5 Representing data categories as terminological entries

7.5.1 General

Many of the data element concepts represented by data categories have long been recognized as embodying many of the reference concepts used in the domain of language resources. It is important to remember that the terms associated with these reference concepts may function differently from the data categories that represent analogous data category concepts in databases. Nevertheless, the basic principles of terminology management provide a well-defined background for describing and expressing concepts in domain specific contexts. Consequently, terminology studies and practice provide valuable principles that can be adopted for the specification and definition of data categories.

One important aspect that such a perspective provides to the representation of data categories is its intrinsic capacity to make a distinction between the two important notions of working language and object language.

7.5.2 Information to be expressed at Description level

The descriptive part of a data category specification is analogous in many ways to a terminological entry as defined in ISO 16642 in that it can be viewed as having a descriptive level (analogous to the terminology entry level), an object language level (analogous to the language section level) and a name section level (analogous to the term section). The descriptive level should be used to record any general descriptive information applicable to the data category. As summarized in figure 6, the following descriptive data categories can thus be associated to this level:

- /entry identifier/ [One and only one]: may only be used when interchange of data category information is made at Description level, in which case it is a mandatory field;
- /definition/ [12620:A.5.1; +]: should be used to provide a reference definition in the registry. As much as possible, the definition should be language and theory neutral. This information is mandatory for each DEC. It may be repeated to provide translations of the definition in other working languages. When necessary, /definition/ may be refined by a /source/ and a /status/;
- /explanation/ [12620:A.5.2; *]: can be used to provide additional information about the data category that would not be relevant for a definition (e.g. more precise linguistic background for the use of the data category);
- /example/ [12620:A.5.4; *]: at TE level, the use of examples should be limited to those that illustrate the data category in general, excluding language specific usages, which should be documented at LS level;
- /source/ [12620:A.10.19; one and only one per definition]: may refine[3] /definition/, /explanation/, or /example/ to indicate the source from which the corresponding text has been borrowed or adapted.

---

[3] Such a refinement can be expressed in GMT (cf. ISO 16642) as follows:

```
<brack>
 <feat type=”definition”>Cas linguistique utilisé …</feat>
 <feat type=”source”>TLFI, nominatif, a.2</feat>
</brack>
```

When a definition is compiled from more then one source, this field can be repeated. The /source/ field should not be used alone at TE level;

- /status/ [not in 12620[4]; one and only one per definition]: may refine /definition/ to indicate approval, acceptability, or applicability in a given context. In particular, it should be used to record alternative definitions, or older definitions that one mat want to keep for documentary purposes. The /status/ field should not be used alone at TE level;

- /profile/ [*]: should be used to relate the current data category to one or several views (e.g. Morpho-syntax, Syntax, Metadata, Language description, etc.), and can thus be iterated;

- /conceptual domain/ [one and only one; ISO/IEC 11179-3]: This field is used to relate the category under description with the set of all its possible values (expressed as a list of data categories). When necessary a datatype (in the sense of XML schemas) may be provided instead of a list of values;

  Ex.: the /conceptual domain/ for /gender/ could be {/masculine/, /feminine/, /neuter/}

- /note/ [12620:A.8; *]: additional information associated with the TE level, excluding technical information that would normally be described within /Explanation/;

- /broader concept generic/ [A.7.2.1; ?]: May be used to point to a more general data category (e.g.: from /Common noun/ to /Noun/; from a given language to a language group or family);

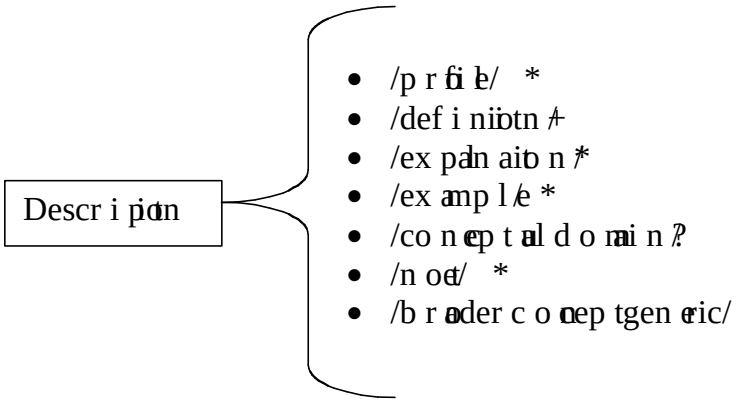

Figure 6 — Data categories associated to the CE level.

7.5.3 Information to be expressed at LS (Language Section) level

In a Descriptive level, the Object Language Section level, which is optional and can be repeated, is intended to describe the information that is specific to a given object language. In simpler cases, it is used to record the way a concept is expressed in a given language. When needed, this shall be used to record all language specific aspects of a data category. The following data categories will thus be used at this level[5]:

- /language/ [12620:A.10.7.1; one and only one per LS]: to identify the language being described (i.e. *object language*, as defined in ISO 16642). In further developments, it may be necessary to refer also to language families;

- /definition/ [*]: to define the data category when it occurs in a specific system within a language, so that it impacts on the accuracy of the reference definition;

---

[4] This data category is inspired from /Term status/.

[5] ISO 12620 references are not repeated here.

- /example/ [*]: provides an example of how the data category is used for the current object language;
- /explanation/ [*]: additional explanation specific to the use of the data category in the object language;
- /source/: see TE level;
- /conceptual domain/ [?; ISO/IEC 11179-3]: to be used when a data category is to be associated to a specific subset of the values declared at TE level (for instance /gender/ would have {/masculine/,/feminine/} as a conceptual domain in French);

Ex.: the /conceptual domain/ for /gender/ in French could be restricted to {/masculine/, /feminine/}

- /note/ [*]: additional information associated with the LS level, excluding technical information that would normally be described within /explanation/;

7.5.4 Information to be expressed at NS (Name Section) level

The Name Section level shall be used to record a possible appellation for the data category in the object language elicited at Language Section level. The Name Section level may be repeated within a Language Section level. The descriptive elements associated with the Name Section level are the following ones:

- /name/ [one and only one per NS]: one word or multi-word unit used to refer to the data category for the corresponding object language as expressed in the encompassing LS block. Names given to a data category shall not be used for the purpose of identifying a data category (see /identifier/).
- /name status/ [inspired from 12620A.2.9 (/term status/); ?]: with the following conceptual domain: {/standardized name/, /preferred name/, /admitted name/, /deprecated name/, /superceded name/} (taken as such from ISO/IEC 11179; to be discussed depending on our need in TC37)

8 Management procedures for a central data category in TC37

8.1 General organization

As shown in figure 7, it is suggested that ISO committee TC37 should implement one central data category registry[6] encompassing all its possible activities in the domain of data representation and coding. The registry is the place where data categories are maintained, whether they represent placeholders in a data structure (i.e. “complex data category”, such as /gender/), or values for them (“simple data category”, e.g. /masculine/). As the details related to the actual representation of data categories will be dealt with in further sections, we focus here on the overall logics of the registry.

Even if centralized, the data category registry is based on the hypothesis that it can be accessed through *thematic views*, i.e. domains of activities, which, in the scope of TC37, requires the identification of specialized subsets of the registry. For instance, such a view may correspond to the data categories that can be used in morpho-syntactic annotation, or also the various data categories involved in language coding.

Not only can do thematic views corresponds to ways of accessing the registry, but they also correspond to a basis for filling it in with new data categories and maintaining them. It is indeed anticipated that the management of the registry should not be fully centralized, but based on a structure that will both put together the right expertise within a subfield of linguistic resources and ensure a good coherence within the registry.

---

[6] Other data category registries may be deployed by specific community for their internal use. Such works can be submitted at a later stage to the ISO TC 37 DCR, as long as a shared interchange format for the representation of data categories is adopted.

Accordingly, the decision process that leads to the introduction or revision of a data category into the registry is organized into two steps:

- A *selection process* by which a *thematic committee* identifies those data categories that are relevant for a certain application field within TC37;
- A *harmonization process*, operated by a *DCR board*, which guaranties the coherence of new proposals with the scope of the registry and data categories it already contains.

In the following sections, a presentation is made of the various actors and bodies involved in the maintenance of the data category registry.

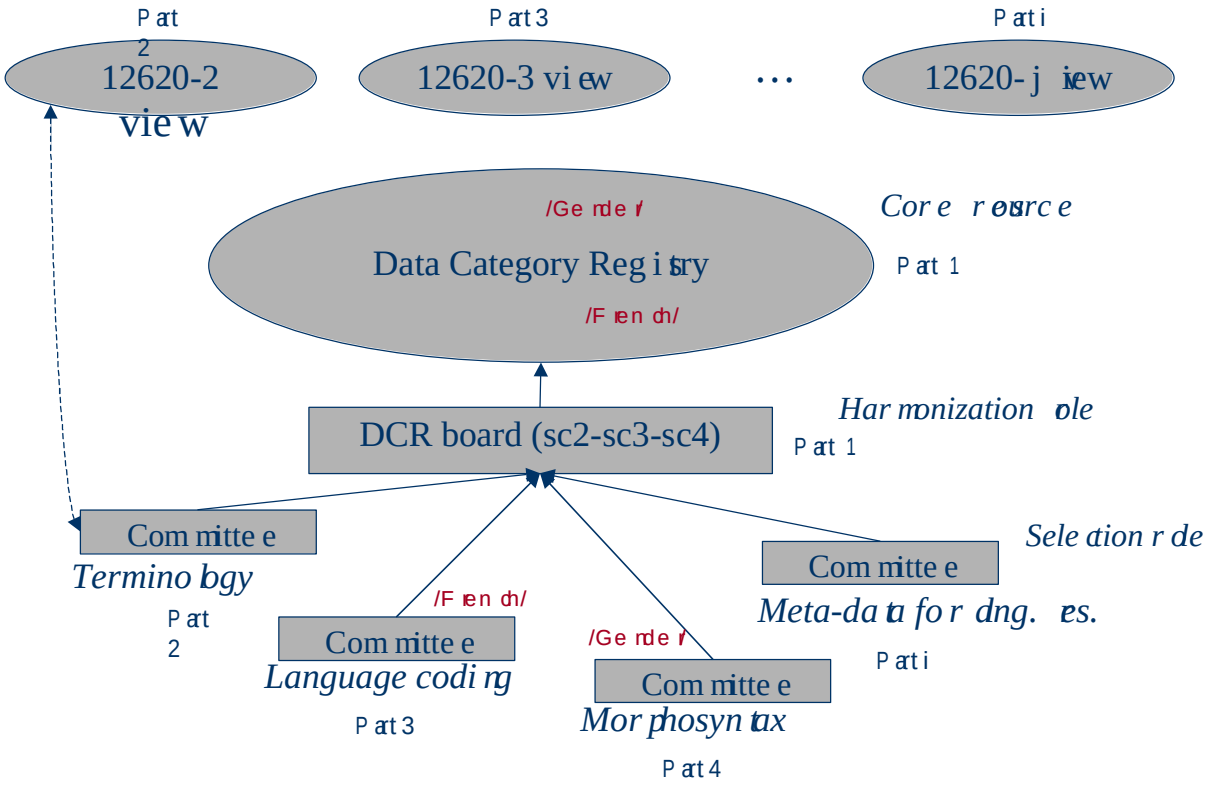

Figure 7 — General organization of the TC37 data category registry.

8.2 Users, experts

As a general rule, the data category registry is freely available on-line for public consultation. This will ensure that language practitioners and implementers in any situation systematically use it.

A specific category of users, named *experts*, may submit proposals for creating or revising a data category. An expert is anyone who has declared himself in the registry, through an on-line form. To prevent spamming and misuse, an expert may be drawn out by the DCR board, when his deeds have clearly been outside the scope of activity of the registry.

8.3 The thematic committees

8.3.1 Constitution

As implied above, different *thematic domains* within TC 37[7] will have a need to arrive at a specific data category selection designed to meet their own data documentation needs. One such thematic domain is terminology management, and the data categories in ISO 12620 represent one such DCS. Another DCS, such as for lexicography or language resource metadata, might contain a subset of data categories already included in 12620 Pt 2 (e.g., administrative and linguistic data categories) as well as contribute a number of new data categories native to its own thematic domain but not generally used in terminology management. In this regard, it is inappropriate to think of subsetting 12620 but rather to view 12620 as it now exists as a subset of the global registry. In order to define any given data category selection, a thematic committee is constituted following a resolution from one of the TC37 sub-committees or directly at TC37 level, either because it is motivated by an on-going standard development, or simply because a new descriptive domain appears to be essential to the SC or TC in question. In all cases, it is necessary for the committee to provide a document (New thematic committee proposal) that states the purpose and scope of the committee and its possible relations to existing committees in the registry, three months prior the SC or TC plenary that will validate the proposal.

By default, the committee is composed as follows:

- a chair, appointed by the SC at the time of the creation of the committee;
- a list of experts designated by the P- and O-members of the SC;
- a group of relevant experts proposed by the chair of the committee, when necessary expertise is to be considered for the good progress of the committee work. The total number of experts designated directly by the chair should not exceed 50% of the total number of experts in the thematic committee;

When created, the committee is also assigned to a view, which is instantiated in the registry.

In some specific cases, for instance when a thematic committee is appointed in conjunction to a standard under development, the can be some variations on the above-mentioned principles. For instance, it could be possible that the existing ISO 639/Ras-JAC, together with its existing procedures and stricture, be seen as being the thematic committee for language description.

8.3.2 Procedure of work

8.3.2.1 Submission of a new data category

The submission of a new data category shall be based on a description compliant with the DCIF model, with the following restrictions:

- only the Description level of the DCIF model shall be documented;
- at least a definition in English shall be provided at Description level and, if applicable, at Language Section level;
- if the definitions are taken from known source, those sources shall be provided;
- at least a profile associated to the data category shall be provided;

[7] In some cases, this activity will be conducted in close collaboration with other ISO committees, such as ISO TC 46/SC 4 for language codes/descriptions; or JTC 1/SC 36 for CALL (Computer Assisted Language Learning) applications.

- at least an English name for the data category shall be provided;
- a note justifying the relevance of the data category to the field of language resources shall be provided.

[Contributions welcome]

8.3.2.2 Modification proposal of a new data category

When a modification is submitted, the following requirements shall be fulfilled:

- the data category shall be uniquely identified (with /identifier/ and /version/);
- only the fields for which a change is suggested (either a modification or a new piece of information) shall be informed;
- a note justifying the relevance of the proposed change shall be provided.

[Contributions welcome]

8.3.2.3 Assignment of an existing category to the view

When a request for assigning a data category to a view is submitted, the following requirements shall be fulfilled:

- the data category shall be uniquely identified (with /identifier/ and /version/);
- a note justifying the relevance of assigning the data category to the profile shall be provided.

[Contributions welcome]

8.3.3 The DCR board[8]

8.3.3.1 Constitution

The DCR board has the duty to ensure that the scope and the coherence of the registry is maintained. It plays a harmonizing role with regards to the proposals that are submitted by the thematic committees.

The DCR board is composed as follows:

- a group of experts designated by the P- and O-members of the SC;
- a chair, appointed by the TC37 plenary for a period of two years, which may be renewed once;

8.3.3.2 Procedure of work

8.3.3.2.1 Validation of a thematic committee proposal

The submitted DC shall be defined by the minimum necessary criteria for a DC as outlined in section X.X.Y (Submission of a new data category)

The submitted DC shall have a status of 'submitted' at DC board level.

---

[8] This could also be called the DCR Registration Authority (or DCR-RA).

DC board level shall ballot the DC.

If positive votes of more than 70% (to be discussed) are received, the status of the DC shall increased to 'board level standard'. If less than 70% is received the DC shall be given a 'rejected' status and reasons for rejection will be fed back to the proposer. Where a rejected DC is proposed again following modification, this shall follow the process for a new DC but shall contain notes regarding the previous submission.

The question would arise here of when a DC can be published in a standard. If 'board level standard' is set, the DCR management could still reject or request modifications to the definition etc if there was a general conflict.

8.3.3.2.2 Publication of a reference version of the registry

Every six months, the registration authority in charge of the maintenance of the DCR, and with the approval of the DCR board shall issue an updated version of the DCR that shall be considered as the official reference for the following period. It shall inform TC 37 Member bodies and liaison as well as SC scretary and chairs of the changes that occured as compared to the previous issue (additions, modifications, deprecations).

8.4 Personal workspace

As an additional feature, the data category registry could provide any expert with the possibility to manage his own workspace, where he could draft his own proposals of new categories or modification of existing ones in the registry. The expert could then publish his work to one or several experts to receive feedback on his proposal, in preparation of an official submission to the relevant thematic committee.

Annex A
(normative)

The GMT DTD to be used for interchange of Data Category Selections

The following GMT DTD implements the DCIF meta-model and shall be used for exchanging data category information. It is based compatible with the GMT format described in ISO 16642 except for the values of the 'type' attribute associated with the 'struct' element. More precisely the various nodes of the DCIF meta-model have been encoded as stated in the following table.

Level name in the DCIF meta-model	*Code to be used in the GMT DTD*
Data Category Registry	DCR
Global Information	GI
Data Category	DC
Description	Desc
Language Section	LS
Name Section	NS
Administration Identification	AI
Administration Record	AR
Registration Group	RG
Submission Group	SubG
Stewardship Group	StewG

```
<!ELEMENT struct ((feat|brack)*, struct*)>
<!ATTLIST struct
      type (DCR|GI|DC|Desc|LS|NS|AI|AR|RG|SubG|StewG) #REQUIRED
      id ID #IMPLIED
      target CDATA #IMPLIED>

<!ELEMENT feat (#PCDATA | annot)*>
<!ATTLIST feat
      type CDATA #REQUIRED
      target CDATA #IMPLIED
      source CDATA #IMPLIED>

<!ELEMENT brack (feat, (feat|brack)+)>
<!ATTLIST brack
      source CDATA #IMPLIED>

<!ELEMENT annot (#PCDATA)>
<!ATTLIST annot
      type CDATA #REQUIRED
      target CDATA #IMPLIED>
```

Annex B
(normative)

Printed representation of a Data Category Selection

This section should be devised in conjunction of the current work on ISO 12620-2 (Data categories for terminology) in order to identify the optimal subset of the DCIF model that should be used for describing a data category selection (DCS). Once decided, this subset shall be used uniformly in any future additional part of the ISO 12620 standard series.

Annex C
(normative)

A DCIF subset for data category information interchange

This section should describe a subset of the full DCIF model allowing the simple exchange of data category information between project or applications. In particular, it should be more flexible as to the precise description of administrative information attached to a data category.

Annex D
(informative)

Example

The following example is an XML representation of the core information associated to a data category on the basis of the principles described above. It is expressed in the GMT format as described in Annex A, and only implements the Description level of the DCIF metamodel.

```
<struct type="Desc">
      <feat type='entry identifier'>Gender</feat>
      <feat type='profile'>morpho-syntax</feat>
      <brack>
            <feat type="definition" xml:lang="fr">Catégorie reposant, selon les
            langues et les systèmes, sur la distinction naturelle entre les sexes
            ou sur des critères formels. Genre naturel, grammatical; genre animé,
            inanimé, genre féminin, masculin, neutre; genre adjectif, substantif
            des deux genres.</feat>
            <feat type="source">www.atilf.inalf.fr Tlfi, GENRE, d, gramm</feat>
      </brack>
      <feat type='example'> </feat>
      <feat type='explanation'>…</feat>
      <feat type='conceptual domain' target='#MASCULINE #FEMININE # NEUTER'/>
      <struct type="LS">
            <feat type='language'>English</feat>
            <brack>
                  <feat type='definition'>…</feat>
                  <feat type='source'>…</feat>
            </brack>
            <feat type='example'>…</feat>
            <feat type='explanation'>…</feat>
            <struct type="NS">
                  <feat type="name">Absolutive</feat>
            </struct>
      </struct>
</struct>
```

Annex E
(informative)

Mapping between DCIF and the Salt format for datacategorie representation

E.1 Introduction

The SALT project (http://www.loria.fr/projets/SALT) proposed a first XML based representation for data categories. Although this format did not contain all the features required or recommended in this document, it has been used in the context of several concrete applications, and in particular as the underlying format for the revision of ISO 12620 part two (Data categories in terminology). The Salt format contains two types of information:

- descriptive attributes used for identifying and documenting a data category;
- supplementary attributes describing the implementation of a data category in a given context, i.e. a data category instance.

E.2 Attributes in the SALT format corresponding to the identification and documentation of a data category

Table E.1 below shows a mapping between the various fields of the SALT RDF format and the DCIF model. In this table, the following notation has been adopted:

- kjuhkdqjhs RDF
- fields in the DCIF format are identified by a path statement combining a) a traversal of the metamodel from the root of the DCIF model (DCR) down to the relevant level and b) the name of the data category attached to this level.

-

Table E.1 — Attributes in the SALT format corresponding to the identification and documentation of a data category

RDF attribute in the Salt format	*DCIF counterpart*	*Comment*
DCIdentifier	DCR.DC.AI.AR./identifier/	
DCName	DCR.DC.CE.LS.NS./name/	DCName was meant to be unique in the SALT format, hence duplicating the role of DCIdentifier
DCDefinition	DCR.DC.CE./definition/	
DCParent	DCR.DC.CE./broader concept generic/	
DCComment	DCR.DC.CE./explanation/	
DCExample	DCR.DC.CE./example/	
DCAdmin		

Bibliography