DCR Style Guidelines

According to ISO 12620, and conventions established during the elaboration of ISO 12620:1999, a data category specification shall conform to the following guidelines:

There shall be at least:

- a camel-case **Identifier**;
- one English **Name** for the data category in the English Language Section, associated with the specification of the **Status** for this name;
- one **Definition** in the English Language Section, together with its **Source**;
- a statement in the **Justification** field indicating the relevance of the data category to the field of language resources;
- a specification of the **Conceptual Domain** for complex data categories.

These rules are not only viewed as best practices for all DCR data category (DC) specifications; they are checked by the DC validator and reported as errors if missing. *It is much easier to add required information when creating an entry than to try to go back to find and correct missing elements later*.

In addition, the following requirements apply for DC specifications that are submitted to the standardization process:

- at least one thematic domain profile value associated with the data category.

1 Names and Identifiers

Data category identifiers and names are subject to certain constraints and best practices.

1.1 Non-mnemonic identifier

A ***persistent identifier (/pid/)*** is automatically assigned by the system when the data category specification is created. It consists of a simple numerical value plus the ISOcat identifiers, e.g., **pid=http://www.isocat.org/datcat/DC-1345.** This identifier is used internally, but also for *w*hen referencing a data category in a bibliographical or technical context from an external resource, such as from an ontology or relation registry. These references often also include the human-readable mnemonic identifier.

1.2 Mnemonic identifier

The ***Identifier*** field in the data category specification comprises the camel-case representation, usually of the most widely used data element name. For instance, *partOfSpeech* is the identifier for the data category */part of speech/.*

XML Naming Rules: The identifier must adhere to the following XML naming rules and conventions:

- Identifiers can contain letters, numbers, and other characters;
- Identifiers shall not start with a number or punctuation character;
- Names Identifiers shall not EVER contain spaces;

- Any string can be used; no words are reserved.

These are XML rules and they are non-negotiable and non-discussable!

Best DCR Identifier Practices

- Identifier names should be descriptive and transparent.
- Choose names that follow best practices for your thematic domain in English.
- Be sure your English spelling is correct – cognate equivalents can be deceiving, e.g., *appellative noun*, not *apelative noun*.
- The Identifier validation function will inform you whether the identifier already exists in the system. The identifier does not have to be unique (because the non-mnemonic identifier is unique), but check to see whether your data element concept has already been entered into the system. You may be able to add your domain to the existing DC profile, but you may also find that you need to add additional information, alter the definition, or change the value domain. In future there will be a feature to allow you to capture the DC in your own working space and propose either a new DC or editing changes to the existing DC.
- Multiword names shall be written with camel case: *bookTitle*.
- Names should be concise, if possible, like this: *bookTitle* not like this: *theTitleOfTheBook*.
- Avoid "-" characters. If you name something *first-name*, some software may think you want to subtract "name" from "first".
- Avoid "." characters. If you name something *first.name*, some software may think that "name" is a property of the object "first", or that .first is an extension of some sort.
- Avoid ":" characters. Colons are reserved to be used for namespaces.
- ISOcat is Unicode compliant. Name Sections and names should be able to accommodate any UTF8 characters, but it is essential that the identifier be XML compliant. If in doubt, check http://www.w3.org/TR/REC-xml-names/#NT-LocalPart.
- You can enter names in other languages in the respective Language Section.

When writing about data category identifiers, the TC 37 community uses the convention */normativeAuthorization/*, */preferredTerm/*, with the data element name itself in camel case and in italics, preceded and followed by forward slashes. The same convention can be used for data element names that are not in camel case: */normative authorization/, /preferred term/.*

1.3 Name Sections in the Language Section

There shall be at least one name in the English Language. It can allow spaces and be more legible, and names may vary considerably depending on the domain being documented. Note that both the Data Element Name section and the Name Section in a given language can be repeated as needed to document different names assigned to the same data category concept in different environments or different communities of practice within the same thematic domain, which means that you may enter multiple names. For instance, for */part of speech/* you might also add */word category/*. Indicate the Status of each name by expanding the pick list and choosing the appropriate status. Here you may find that the preferred name for your domain is different from that in another domain.

1.4 Data Element Names

Language independent but application domain-specific names can be entered as Data Element Names. This may be the same name as that used for the identifier or a name used in a language section, but perhaps not. For instance, current */LexTermType/* as an identifier may be */term type/* in terminology applications. Or ADJ may be in use as an abbreviation for /*adjective*/ in a specific tagset. The data element name does not have to written in camel case in any event. It is strongly suggested that data element names be represented in lowercase, but no other conventions are required. The source of the data element name, e.g., the application, should always be provided.

2 Definitions

Definitions shall follow the rules outlined for intentional definitions in ISO 704 unless otherwise indicated:

- They should consist of a single sentence fragment;
- They should begin with the *superordinate concept*, either immediately above or at a higher level of the data category concept being defined;
- They should list critical and delimiting *characteristic(s*) that distinguish the *concept* from other related *concepts*.
- Note: Actual concept systems, such as are implied here by the reference to broader and related concepts, should be modeled in Relation Registries outside the DCR. Furthermore, different domains and communities of practice may differ in their choice of the immediate broader concept, depending upon any given ontological perspective. Harmonized definitions for shared DCs should attempt to choose generic references insofar as possible.

Example: **grammatical gender**

DEFINITION: A grammatical category that indicates relationships of form and agreement between nouns, pronouns, and adjectives in a given discourse segment.
NOTE: The concept of gender varies from language to language and is not a universal feature of all languages. It is false to assume that grammatical and natural gender can be conflated – in many languages gender is dependent on word form, etymology, or other factors rather than the natural “sex” of the object or individual in question.

EXAMPLE: In French, *vie* (life) is feminine and is used with a feminine article such as *la*, the feminine pronoun *elle*, and feminine adjective endings, e.g., *une longue vie*.

Note that there is no finite, independent verb in the definition: it comprises a predicate consisting of the implied subject “grammatical gender” followed by the unexpressed copula [is] and the implied predicate noun, the superordinate concept “a grammatical category,” which is then qualified by the characteristics “that indicates grammatical relationships” and “between words in sentences”.

Definitions should NOT be tautological with respect to the data category name.

Negative example: **zuInclusion**

DEFINITION: Inclusion of zu.

Improved definition:
DEFINITION: In German, the insertion of the particle "zu" between a separable prefix and the base verb in order to form the infinitive, in contrast to the standard position of the particle immediately before the infinitive.

EXAMPLE:
übersetzen, meaning *to translate*: I read the text in order to translate it. Ich habe den Text gelesen, um ihn zu übersetzen. (standard position)
übersetzen, meaning *to transport to the opposite side* of something: I came to the river in order to transfer the boat to the opposite side. Ich bin an den Fluß gekommen, um das Boot auf die andere Seite überzusetzen. („zu" inclusion)

Data category specification presentational features:

Data category definitions as represented in the DCR follow agreed presentational features designed to differentiate them from terminological definitions found in the ISO Concept Database and in terminology standards such as ISO 1087. In this regard, they begin with a capital letter, they use articles (*a, an, the*, or in some cases, *any*), and they end with a period.

Any supplemental information should be added in a note or in the explanation class.

3 Source

Any quoted material, such as definitions, notes, and explanations, shall be properly documented in the *Source* class. Original material should be attributed to the individual or group that created it. This applies also in the Language Sections. ***Missing Sources are the major cause for validation failure when running the DC validator*!**

4 Justification

The Justification field is obligatory. Please write a simple statement justifying the relevance of the data category to the field of language resources. You as an expert in your specific domain may feel that the inclusion of the item is self-evident, but if you ever declare your DC to be public, or if you submit it for standardization, a broader group will be looking at it. Especially for standardization, remember that the DCR Board must make the final decision whether to include an item, and some members may not be familiar with all areas of interest. If the utility of a proposal is unclear, it may be rejected.

For instance, */have/* and */is/* are included in the Morphosyntax profile. It is possible to think of two or three scenarios where these items might serve as simple DCs in a markup environment, but it is not clear which of these possibilities applies here. An explanatory comment is needed to clarify their inclusion.

5 Thematic Domain Profile Values

The standardization area of the DCR is divided into Thematic Domains, which reflect particular areas of interest. For more information on Thematic Domains, see http://www.isocat.org/. A data category specification can be created without declaring a value for the Thematic Domain Profile. In this case, the value is set by default to *Private*, which indicates that the data category specification "belongs" to an individual rather than to a thematic domain group (TDG). These data category specifications may remain private or be shared with individuals or made public.

However, if the creator chooses to submit the data category specification for standardization, it must be associated with a TDG by entering the appropriate value in the Thematic Domain Profile. This is important because the TDGs function as the Maintenance Team specified by ISO for the maintenance of standards as databases. If the data category represents a thematic profile that is not yet represented by a TDG in the DCR structure, its creators can request the creation of a new TDG. Such requests are treated as Change Requests and shall be evaluated by the respective Sub Committee (SC) within TC 37, or possibly at the TC level, and if approved as per ISO 12620:2009, the TDG will be officially introduced into the program of the DCR and appropriate experts will be assigned.

6 Other information

Mandatory administration information is generated automatically by the DCR software.

Optional Information, such as additional names and definitions in other languages, should, if applicable and available, be documented. Names and definitions in other languages may or may not be translations of the English forms, depending on common usage. They should, however, follow essentially the same rules as those that apply for English, with the exception that there may be specific rules in some languages that differ from the English language rules. In such cases, care should be taken to adopt meaningful rules for presentation in other languages