

The OWL and the ISOcat: Modeling Relations in and around the DCR

Sue Ellen Wright¹, Marc Kemps-Snijders², Menzo Windhouwer²,

¹Kent State University Institute for Applied Linguistics

109 Satterfield Hall, Kent, Ohio 44242 USA

²Max Planck Institute for Psycholinguistics

Wundtlaan 1, Nijmegen, The Netherlands

E-mail: sellenwright@gmail.com, Marc.Kemps-Snijders@mpi.nl, Menzo.Windhouwer@mpi.nlm

Abstract

The TC 37 Data Category Registry (DCR) provides a web-based environment for specifying data categories (DC) used in language resources. DC specifications comprise a set of administrative and linguistic attributes, such as definitions, conceptual domain specifications, examples, and any DC names used for data category concepts, as well as language-specific versions of names and definitions. In order to avoid a proliferation of relations and ontological hierarchies, the system does not represent relations between DCs. The plan is to build Relation Registries (RR) “in the environment” of the DCR which reference individual specifications via persistent identifiers (PID) and establish meaningful relations between DCs that can be used to build broader language and knowledge resources. This architecture focuses on RDF solutions in general, with a leaning toward OWL DL. Current issues include establishing best-practice approaches for utilizing OWL elements for maximum effectiveness, methods for anchoring DCR-PIDs in relation assertions, and methods for linking to other authoritative resources, such as the ISO Concept Database (CDB).

1. Conceptual Ordering and the DCR

The current TC 37 Data Category Registry (DCR) is based on ISO 12620:2009, which establishes a data model and guidelines for implementing specifications for data categories (DC) used in language resources, including both object names and values, as well as tagging values used in annotation frameworks. Actual data category information is documented in the web-based ISOcat registry (ISOcat, 2010). The predecessor standard, ISO 12620:1999 was limited to hardcopy representation of DCs used in terminology resources. The lack of electronically processable open access to these DCs motivated developers to move the collection to a web-based resource, first in the form of the Syntax pilot registry (Ide and Romary, 2004 and Wright, 2004), and now ISOcat (Kemps-Snijders et al., 2009).

The original 12620:1999 featured a conceptual ordering system that grouped terminological DCs in “logical” sections and sub-categories according to a consensus-based hierarchy. In retrospect, however, it has become clear that this system is not universally acceptable. Diverse groups even within a single thematic domain (e.g., terminology) rarely have the same needs with respect to ordering principles. This potential for a multiplicity of views and overall complexity has been exacerbated over time by the addition of new working groups and projects in TC 37.

Furthermore, the initial standard used the cited ordered list to form semi-mnemonic identifiers, which proves almost impossible to maintain if the set is expanded at any point. Many of these identifiers remain in place in environments such as the TBX standard (ISO 30042:2008), but they are now treated as virtual non-mnemonics. They are not, however, persistent identifiers in the sense discussed in section 4 of this paper (see also ISO DIS 24619 and Windhouwer et al. in these Proceedings).

The Syntax project revealed several principles with regard to ordering systems within a data category registry that are

conformant with best practices in the metadata field in general:

- data category collections should not impose any specific ordering system because the needs and views of different communities and sub-communities of practice using the registry will vary;
- the potential for a multiplicity of approaches, particularly in OWL modeling environments, would place an unnecessary burden on data administration if an effort were made to accommodate complex structural modeling within the DCR itself.

2. Limited ordering features

Despite the prohibition on concept modeling within ISOcat, there are a few structural features in place that do provide for very shallow ordering within the collection.

2.1 Thematic Domain Groups and Data Category Selections

A DC can be entered independently by an individual without assigning it to any particular set, but DCs in general, and specifically any proposed for standardization as per ISO 12620:2009, are assigned to so-called Thematic Domain profiles, such as *Terminology*, *Morphosyntax*, linguistic *Metadata*, etc. (Kemps-Snijders et al., 2008 and 2009, 243). A DC can be assigned to more than one profile, and the full set for any given profile is viewed as a Data Category Selection (DCS) (ISO 12620, Section 4.2). Individuals and working groups can also define subsets of TDG-related profiles for specific purposes, which are also identified as DCSs, and any of these DCSs can also be standardized if desired. For instance, the *Terminology* DCS can be further subsetted into the TBX-Default DCS or a smaller, TBX-Basic subset. Hence there is an implicit simple relation between any DC and the TDG and/or DCS or profile to which it is assigned. These DCs can be displayed as lists associated with the respective profiles or

DCs, but they cannot be further articulated in any kind of hierarchy. Figure 1 illustrates how the *Morphosyntax* TDG has been subsetted in order to facilitate finding specific DCs.



Figure 1 Subset DCs in the Morphosyntax TDG

2.2 Closed DCs and their value domains

Data categories are further differentiated as either complex or simple DCs. The former are associated with conceptual domains, which is to say that they *have* values, and while the latter *are* values and hence do not have a conceptual domain. The current DCR data model supports three types of conceptual domains: *open*, *constrained* and *closed*. Open complex DCs can have any value that conforms to the definition of the concept underlying that particular DC, and these values cannot be predicted, for instance, */term/* can be any word, phrase, or even text chunk that can in a given context be construed as a term.

In contrast, the closed DC */grammatical gender/* as defined in ISO 12620:1999 specifies */masculine/*, */feminine/* and */neuter/* as possible values, with the option for */other/*, as a hedge against possible unknown instances. For instance, although romance languages are limited to */feminine/* and */masculine/* genders, and some languages like German and Russian add */neuter/*, Polish breaks out masculine into */personal masculine/*, */animate masculine/* and */inanimate masculine/*. Less familiar languages could exhibit further variation, which suggests yet another type of conceptual domain, for something like “partially closed” DCs, which would account for the insertion of additional, previously non-enumerated variants.

In contrast to enumeration, constrained DCs are governed by some sort of schema-based constraint, such might be used to restrict values of a */date/* field to a certain range or an */email/* field, which can be restricted using a regular expression specifying it must contain an @ sign.

2.3 *IsA* relations between simple DCs

Despite the prohibition on hierarchies within the DCR, the system does provide for expressing limited *IsA* relations between subsets of simple DCs and their broader (less granular) pseudo-siblings within a long flat list of simple DCs. For instance, in the

Terminology TDG, the complex closed DC */term type/* lists at least 26 different values in its conceptual domain, and can be expanded to include a few more if Japanese script forms are treated as orthographic variants. Using the *IsA* feature, this long unwieldy group can be divided into meaningful subsets, i.e., the values */abbreviation/*, */acronym/*, and */initialism/* are all more granular instances of */abbreviated form/*. Figure 2 illustrates how this one subset can be displayed in the ISOcat interface, and such subsets can also be viewed in the main display pane using the “hierarchical relationships” view. However, it is not possible to display relations between complex DCs or between DCs and containers like those in TBX. This is only possible outside the DCR in some other format like OWL, with its limitless representational range.

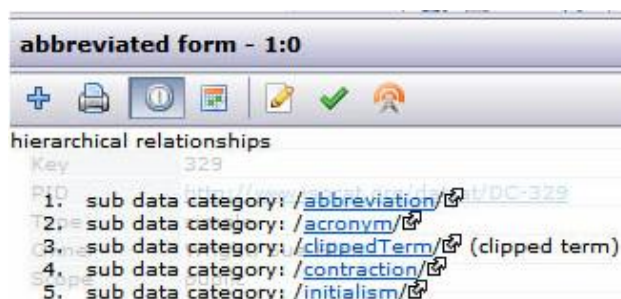


Figure 2 */abbreviated form/* subset of */term type/*

3. Modeling variance

The list of term type values and the ordering principles described in the previous section reflect a terminological approach linked to several standards (ISO 16642:2003, ISO 12200:1999, and ISO 30042:2008). The value domain in the standards (and in ISOcat) is long partly because */term type/* is not only a DC in ISOcat – it is also a metadata category in TBX. Furthermore, creators of terminology management applications – at the behest of impatient users – tend to be parsimonious with GUI real estate. Rarely does any application actually make use of more than a handful of the items in the long list, so individual DCs that might reflect specific data models will exhibit different modelling features.

Not only will different terminological approaches present different data models, and consequently different sets of relations between data categories. Viewed from the standpoint of the Data Category Registry (DCR) as a whole, views on the data will also vary considerably depending upon the needs and perspective of specific thematic domains. Figure 3, for instance, illustrates a common modelling facet viewing the DC */noun/* as a feature perhaps in a lexical or syntactical resource, whereby */noun/* is represented as having values such as */grammatical gender/*, */grammatical number/*, etc.

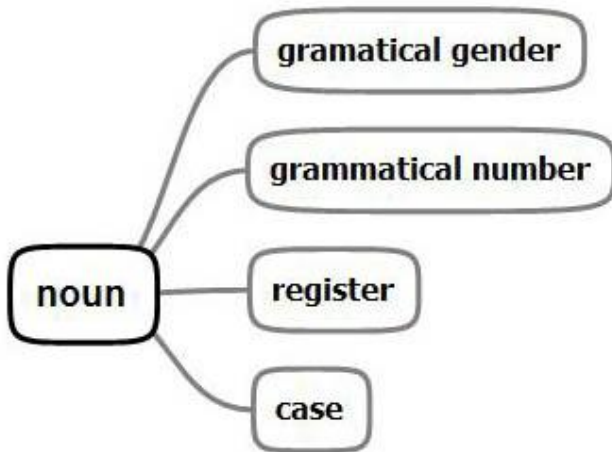


Figure 3 /noun/ modeled for grammar

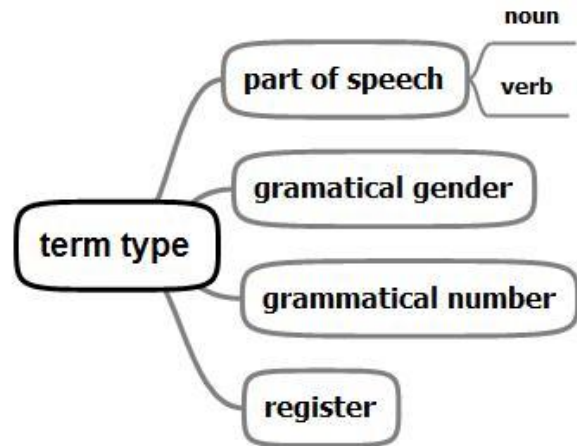


Figure 4 /noun/ as modeled in TBX

From the viewpoint of a terminologist modelling in the spirit of ISO 16642 or TBX, however, /noun/ is itself a terminal value of the data category /part of speech/, while the other attributes that can be associated with a particular term are arrayed in parallel with /part of speech/ as attribute/value pairs embedded in a term information group (see Figure 4). These graphical representations demonstrate a markedly different graphical perspective on the DCs reflecting data modelling variance between communities of practice. It is this kind of situation that has inspired the position maintained by the DCR designers that the kinds of relations that will be asserted with regard to the DCs are highly unpredictable and subject to very different representational needs and modelling environments.

4. Relation registries and persistent identifiers

The DCR development group has characterized such external relation stores as *relation registries* (RR; see Kemps-Snijders et al., 2008). The concept is that users can create their own registries in OWL or SKOS or other RDF representations that will be free to assert whatever relations are needed in the framework of a given environment. Interpretation and use of these relations thus falls within that modelling framework and outside the scope of the DCR. The critical factor here is that the subjects of such assertions must be linkable to their corresponding DC specifications documented in ISOcat via links conformant with persistent identifiers as specified in ISO DIS 24619. For this purpose, ISOcat provides an export utility that expresses any selected DCS as basic RDF descriptions, including information on the data category name (expressed as an OWL label), the internal ISOcat identifier (expressed as a “cool URI), and the data category definition (included in an rdfs:comment):

```

<?xml version="1.0"?>
<rdf:RDF ... [resource declarations] ... >
<rdf:Description
rdf:about="http://www.isocat.org/datcat/DC-329">
  <dcr:datcat
rdf:resource="http://www.isocat.org/datcat/DC-329"/>
  <rdfs:label
xml:lang="en">abbreviated form
</rdfs:label>
  <rdfs:comment xml:lang="en">A term or
lexeme resulting from the omission of any
part of the full term or lexeme while
designating the same concept.
</rdfs:comment>
</rdf:Description>
</rdf:RDF>
  
```

ISOcat RDF output for any given DCS can be patched directly into an RDF resource, where it presents in an RDF editor as a list RDF instances. The current representation of the TBX structure, which began as an OWL DL file, reverts to OWL Full with the addition of these values. This condition explains the use of the OWL:SameAs notation in Figure 5 – given the shift to OWL Full, there is no particular reason to shy away from OWL:SameAs. Nevertheless, one goal is to work out the representation so that output can be efficiently utilized within OWL DL.

5. Outlook

Major goals for 2010 involve the creation of a kind of test bed for relation registries in the environment of the DCR and to further refine OWL representation using DCR RDF output. PID-related efforts are reflected in planned linkage with the ISO Concept Database (CDB) using the following proposed format for the language codes: <http://cdb.iso.org/lg/CDB-00130302-001> (Hebrew).

6. References and abbreviations

CDB – ISO Concept Database, *see* <http://cdb.iso.org/>

DCR – Data Category Registry.

DCS – Data Category Selection.

DIS – Draft International Standard.

Ide, N. and Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *Proceedings of the IVth LREC International Conference on Language Resources and Evaluation*, Lisboa, Portugal, pp.135-138.

ISO – International Organization for Standardization

ISO standards: Geneva, International Organization for Standardization:

ISO 639-1:2002. Codes for the representation of names of languages – Part 1: Alpha-2 Code.

ISO 639-2:1998. Code for the representation of languages – Part 2: Alpha-3 Code.

ISO 12620:1999. Computer applications in Terminology — Data categories

ISO 12620:2009. Terminology and other language and content resources – Data Categories – Specification of data categories and management of a Data Category Registry for language resources

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

ISO DIS 24613:2008. Lexical Markup Framework (LMF)

ISO DIS 24619:2009. Language Resource Management — Persistent Identification and Access in Language

Technology Applications

ISO 32042:2008. Systems to manage terminology, knowledge, and content – TermBase eXchange (TBX).

ISocat. (2010). Data Category Registry: Defining widely accepted linguistic concepts. <http://www.isocat.org/>.

Kemps-Snijders, M., Windhouwer, M.A., and Wright, S.E. (2008). *Putting data categories in their semantic context*. In: *Proceedings of the IEEE e-Humanities Workshop (e-Humanities)*. Indianapolis, Indiana, USA, December 10, 2008.

Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., and Wright, S.E. (2009). ISocat: Remodeling Metadata for Language Resources. In: *Special Issue: Open Forum on Metadata Registries of the International Journal of Metadata, Semantics and Ontologies (IJMSO)*, 4(4), pp 261-276, November 2009.

LMF – Lexical Markup Framework.

OWL Web Ontology Language Overview.
<http://www.w3.org/TR/owl-features/>.

OMG – Object Management Group.

RDF – Resource Description Framework.

SKOS – Simple Knowledge Organization System

TBX – Termbase eXchange, *see* ISO 32042.

UML – Unified Modeling Language

URI – Uniform Resource Identifier.

Wright, S.E. (2004). A global data category registry for interoperable language resources. In: *Proceedings of the IVth LREC International Conference on Language Resources and Evaluation*. Lisboa, Portugal, pp. 123-126.

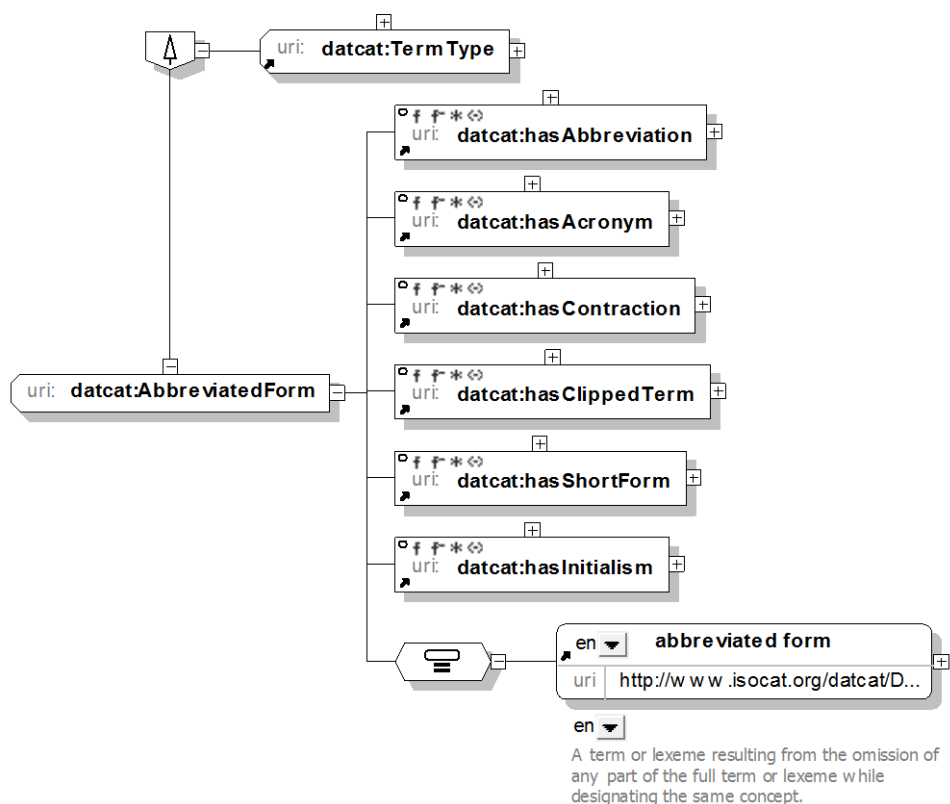


Figure 5: RDF representation with ISocat PID:
/abbreviated form/ **IsA** /term type/; ... **has** [the stated] properties;
... **isEqual to** <http://www.isocat.org/datcat/DC-329>

