Collaboratively Defining Widely Accepted Linguistic Data Categories in the ISOcat Data Category Registry

Menzo Windhouwer[1], Ineke Schuurman[2], Sue Ellen Wright[3]

[1] The Language Archive - DANS
menzo.windhouwer@dans.knaw.nl
[2] Utrecht University, KU Leuven
ineke@ccl.kuleuven.be
[3] Kent State University
sellenwright@gmail.com

Abstract. The ISOcat Data Category Registry (www.isocat.org) has been developed by ISO TC 37 and CLARIN to share and explicitate semantics of data categories used within the linguistic community. Semantics in this large and diverse community are constantly evolving and sometimes conflicting. The ISOcat open registry allows community members to collaborate in defining the semantics of linguistic data categories. The aim is to create a core of possibly officially standardized, well specified and widely accepted linguistic data categories. This demonstration will show ISOcat's features to support direct and indirect collaboration, its efforts to create a set of core data categories for various communities, and possible solutions for current bottlenecks.

Keywords: semantics, collaboration, data category registry, linguistics

1 Introduction

For several decades the ISO Technical Committee 37 (ISO TC 37), *Terminology and other language and content resources,* has worked on standardizing data categories. Data categories (DCs) are defined as "the result of the specification of a specific data field" or "an elementary descriptor in a linguistic structure or an annotation schema" [1]. Examples include */grammatical gender/* and */part of speech/*. A DC specification gives a specific representation, e.g., a data type, of a data element concept. As these data element concepts can be part of an ontology or taxonomy any linguistic resource that reuses known data categories can potentially become part of a semantic network.

By the 1980s the terminology community began defining and sharing frequently used DCs with their abbreviations and definitions. These initial efforts culminated in the ISO 12620:1999 *Data categories* standard [2]. However, new demands and insights quickly showed the limitations of such a rigid paper standard, suggesting that a registry where existing DCs could be managed and new ones easily added was far more appropriate. In 2009 ISO published a revised ISO 12620 [1], which specified a data model for a Data Category Registry (DCR) and procedures to standardize DCs stored in the registry. ISOcat, as developed and hosted by The Language Archive at

adfa, p. 1, 2011.

the Max Planck Institute for Psycholinguistics, implements this data model and supports the specified procedures. While ISO 12620:1999 focused on DCs needed by the terminology community, the ISOcat DCR was adopted by a broader community, especially the European CLARIN infrastructure [3], which introduces additional domains into the registry. [4] describes the history of DCs and DC registries in more depth. In this demonstration we will show ISOcat features that support both ISO TC 37 and CLARIN by involving their communities to collaboratively define (and standardize) DCs and their underlying concepts in ISOcat. However, as the notion of "data categories" is not a common topic in the semantic web discussions, the remainder of this introduction describes the role ISOcat DCs can play in this context.

ISOcat comes from a different tradition than the semantic web and linked data. ISO TC 37 tends to develop dedicated UML meta models, *e.g.*, the recent Lexical Markup Framework (LMF). Such meta models are instantiated in an application specific model using specific selections of ISOcat DCs. The application specific models link their DC instantiations to their counterparts in ISOcat via 'cool URI's', which are uniquely assigned to each individual DC. These DCs can thus very naturally also be used by RDF-based resources and knowledge bases to support semantic mapping.

```
@prefix dcr: <http://isocat.org/ns/dcr.rdf#> .
...
:partOfSpeech a owl:ObjectProperty ;
  dcr:datcat <http://isocat.org/datcat/DC-1345> ;
  rdfs:label "part of speech"@en ;
  rdfs:comment "Term used to describe how a particular
  word is used in a sentence."@en .
...
```

[5] provides guidelines on how to annotate RDF-based resources with references to ISOcat DCs to indicate shared semantics, as shown in the example above. Notice that the dcr:datcat predicate is an annotation property, which reflects the current approach to annotate existing resources and schemas with ISOcat DC references. However, as stated earlier, the underlying DC concepts may appear in an ontology or taxonomy. Currently a companion registry called RELcat [6] is under construction to allow the specification of community- or even user-specific ontologies, *e.g.*, OWL-based, and taxonomies, *e.g.*, SKOS-based, on top of the ISOcat DCs.

2 ISOcat: a collaborative space to define linguistic concepts

The global set of DCs in use by a group as diverse as the linguistic community will never be stable. Assessing whether a new DC should be created or an old one adapted is not something that can be fully automated. ISOcat is thus an open registry where, when needed, every user can create her own DCs. However, the aim is still to establish a stable set of core DCs to be used by a majority of the community for common tasks. This core set can adapt over time when visions and needs shift, *e.g.*, due to new theories or technologies. The registry should help a user to select DCs for a specific language resource type, assist user collaboration, and even support data interchange or interactivity across applications. To illustrate the various collaboration features, various usage levels are sketched in the remainder of this section.

Guest access: Anyone can access the ISOcat DCR to search and browse in the public workspace. This workspace contains DCs and selections (groups of DCs) made public by users and ISO TC 37 related expert groups for domains like Metadata and Morphosyntax. The user interface provides a ‘basket’ in which users can collect interesting DCs and save them to their own machines in various formats.

Private workspace: If just viewing the public part of the registry is the only goal, guest access to ISOcat might be enough, but to really work on a project to annotate a linguistic resource type with DCs, the user should register herself in ISOcat. A registered user can use the ‘basket’ to create a persistent selection which can later be re-used. Also new private DCs can be created and later edited.

Shared workspace: The collaborative features of ISOcat support work in larger teams. A user can create a user group and invite other users to become members. These members can share selections and data categories. Groups can also set up a private forum to discuss issues involving data categories and selections.

Public workspace: Once a user or a group is satisfied with a selection containing public and/or their private data categories for a resource type, they can make it public, *i.e.,* all users, including guests, of ISOcat can then see and refer to these data categories. The group can also start up an additional public forum or make their existing private forum public. Registered users can always be contacted via a mediated email.

DCs and selections in the public workspace can be assessed and used by everyone, but the quality and consistency of the DC specifications can vary widely. Also, due to the open nature of the registry, doublettes can easily be created. This can make it hard for users to select a specific DC among various candidates. The core of standardized DCs should help, *i.e.*, if a candidate is a standardized DC, the user should select that one. The next section describes the collaborative standardization process.

3 Standardization by ISO Technical Committee 37

ISO TC 37 has established a wide range of Thematic Domain Groups (TDGs). Each one has a chair, a number of experts selected by ISO TC 37 member countries and additional experts invited by the chair. Initially these groups will standardize existing sets of DCs, *e.g.*, the ISO 12620:1999 DCs in use by the TermBase eXchange standard for Terminology and a set of Metadata DCs based on existing metadata element sets. These standardized selections then will form a coherent core reflecting the current state of the art in their domains. However, as mentioned before, these domains will be continually in flux. Thus users can submit additional data categories or request changes to existing DCs. ISO 12620:2009 provides the procedures to officially standardize these submissions. The workflow of this procedure is shown in Fig. 1.

A submission is made by one or more users, *i.e.*, by forming a submission group, to a specific TDG. The experts in this group evaluate the DC specification or change request, discuss it, in case of a change request implement it, and in the end conduct

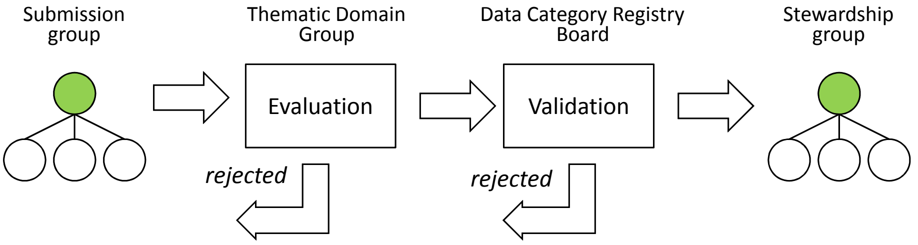

Fig. 1. ISO data category standardization process

a ballot. The result of the ballot leads to either acceptance, in case of at least 70% positive votes, or rejection. Rejected DCs are returned to their original owners. Accepted DCs go on to the validation phase. In this phase the DCR Board, which is another group of experts from a wide range of domains mostly selected by the member countries, has another look at the DCs. While a TDG focusses on a specific domain, the DCR Board looks especially at harmonization issues between various thematic domains. This step also ends with a ballot, where accepted DCs acquire standardized status, while rejected DCs are returned to the TDG, where the chair decides on follow up action. Standardized DCs are no longer owned by an individual ISOcat user but by the TDG, which assigns a stewardship group to maintain the DC specification.

This process is thus quite clearly described and has been implemented in ISOcat. It is also a very collaborative process involving a number of expert groups which have to discuss and in the end make decisions through a ballot. The original aim of ISOcat has been to make the bulk of this process public, *e.g.*, each submission has an associated public forum where the experts discuss the issues and other users can also get involved. However, this process inside an open registry is a significant break with the more traditional paper process common to ISO standardization, and the ISO TC 37 community of experts has not successfully transferred the standardization process into ISOcat. The upshot of this situation is that there are still no standardized DCs in ISOcat. A redesign of the standardization process is thus at hand, including most likely a return to the more traditional ISO process by linking standardization of DCs and/or selections to the standardization of specific meta models.

The continued absence of standardized DCs for many domains in the registry poses problems for adaptation by potential users, *e.g.*, how widely accepted are all these DCs owned by other users? Standardization indicates peer review by experts and until now a clear indication of this process has been lacking. The next section describes how ISOcat has been extended to create an intermediate recommendation layer between the private user workspace and the envisioned standardized core.

4 Recommendations by CLARIN-NL/VL

The large scale European infrastructure project CLARIN has adopted ISOcat, initially especially for its Component MetaData Infrastructure (CMDI) [7]. In CMDI metadata components, elements and values can be associated with a DC, which allows sharing semantics even though different terminology and structures are used. Just like the ISOcat DCR, the CMDI Component Registry is open and any user can create metadata profiles and components. While doing so, the user interacts with ISOcat to select DCs from the Metadata thematic domain, or when needed, creates a new one. The Dutch and Flemish CLARIN national projects (CLARIN-NL/VL) would like to assist the user in selecting and/or creating high quality DCs. Quality is especially important as CLARIN-NL/VL has adopted ISOcat for use by language resources in general, not only for their metadata. For this reason they have assigned an ISOcat coordinator, who is basically the chair of the CLARIN-NL/VL ISOcat group. CLARIN projects share their DCs and selections with this ISOcat group, which allows all the group members to review and reuse them. The coordinator plays a prominent role here by providing training on the use of ISOcat, guidelines on the features of a proper DC specification and feedback. A CLARIN-NL/VL private ISOcat forum is used to coordinate these activities. To make the results of this peer review process visible she, representing a sizeable group of ISOcat users, can mark DCs as “recommended by CLARIN-NL/VL”. This recommendation can thus help users to select a proper DC. It is important to note that this new recommendation feature is generic, *i.e.*, it is possible for other ISOcat groups to also recommend DCs.

In addition to the new recommendations option the existing group support forms the basis for a community-specific view. In such a view the ISOcat user interface only shows the DCs and selections shared with and made public by a specific group, so the CLARIN-NL/VL ISOcat view only shows selective parts of the DCR deemed relevant for that community. The view is easily adapted by adding or deleting DCs and/or selection from the shared workspace.

5 Conclusions and future work

This demonstration has shown that ISOcat offers many ways for users to collaborate. The aim is to share and explicitate semantics, which is important to ensure future interpretation, of language resources, *e.g.*, lexica or detailed transcriptions of endangered languages. But already now the semantic network, which can be built on top of the existing structure, can be exploited for resource discovery. ISOcat has been successful in acquiring an active user base (at time of writing there are around 500 registered users), *i.e.*, due to its origin in ISO TC 37, the associated ties to ISO standards and its adoption by CLARIN. Nevertheless, the coherent core of qualitative DC specifications has not yet emerged from the sometimes confusing mass of DCs (at time of writing there are around 5,000 data categories).

Current development focuses on continuous improvement of ISOcat’s collaborative features and keeping close touch with the user communities in various ways in order to arrive at such a core or, in all likelihood, multiple domain-specific cores. Another focus is extending the semantic network by working on companion registries, *e.g.*, RELcat, which will bring the DC-based approach considerably closer to the semantic web.

References

1. ISO 12620, *Specification of data categories and management of a Data Category Registry for language resources*, ISO, 2009.
2. ISO 12620, *Data categories*, ISO, 1999.T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne, K. Koskenniemi. *CLARIN: Common Language Resources and Technology Infrastructure*. Sixth International Conference on Language Resources and Evaluation (LREC'08). Morocco, May 28-30, 2008.
3. S.E. Wright, M. Windhouwer, I. Schuurman, M. Kemps-Snijders. Community efforts around the ISOcat Data Category Registry. In I. Gurevych, J. Kim (eds), *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer. 2013.
4. M. Windhouwer, S.E. Wright. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff, S. Hellmann (eds), *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata* (LDL 2012), Germany, 2012.
5. M. Windhouwer. *RELcat: a Relation Registry for ISOcat data categories*. Eight International Conference on Language Resources and Evaluation (LREC’12). Turkey, May 23-25, 2012.
6. D. Broeder, O. Schonefeld, T. Trippel, D. v. Uytvanck, A. Witt. *A pragmatic approach to XML interoperability – the Component Metadata Infrastructure*. Balisage. Canada, 2011.