Experiences with the ISOcat Data Category Registry

Broeder, Daan; Schuurman, Ineke; Windhouwer, Menzo

published in
Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)
2014

document version
Publisher's PDF, also known as Version of record

Experiences with the ISOcat Data Category Registry

Daan Broeder[1], Ineke Schuurman[2], Menzo Windhouwer[3]
[1] The Language Archive, MPI for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[2] Utrecht University, KU Leuven
Blijde Inkomststraat 13, B-3000 Leuven, Belgium
[3] The Language Archive, DANS
Anna van Saksenlaan 51, 2593 HW The Hague, The Netherlands
Daan.Broeder@mpi.nl, ineke@ccl.kuleuven.be, Menzo.Windhouwer@dans.knaw.nl

Abstract

The ISOcat Data Category Registry has been a joint project of both ISO TC 37 and the European CLARIN infrastructure. In this paper the experiences of using ISOcat in CLARIN are described and evaluated. This evaluation clarifies the requirements of CLARIN with regard to a semantic registry to support its semantic interoperability needs. A simpler model based on concepts instead of data categories and a simpler workflow based on community recommendations will address these needs better and offer the required flexibility.

Keywords: open registries, community involvement, standardization

1. Introduction

This paper describes the experiences with the ISOcat Data Category Registry (DCR, www.isocat.org) of the Dutch and Flemish national CLARIN initiatives (www.clarin.nl, and within ISOcat known as the CLARIN-NL/VL group), who has been a forerunner in the European-wide CLARIN infrastructure (www.clarin.eu) in this area. These experiences are also valuable for other communities using the DCR or even other types of semantic registries.

One of the aims of the European CLARIN infrastructure is to allow scholars to easily find and integrate data from a wide range of sources. This brings not only the problem of a broad diversity of formats and data structures, but also of terminology and semantics. Semantic interoperability problems are not new and CLARIN worked closely together with the ISO Technical Committee for *Terminology and other language and content resources* (ISO TC 37) on the use and further development of the ISOcat DCR.

CLARIN needs to provide for a broad linguistic community and considerable effort was spent on making ISOcat both in technical and organizational aspects more suitable for that community. Within the CLARIN infrastructure two types of data can be distinguished: (1) the linguistic resources as archived by the CLARIN centers, and (2) the metadata about these resources as offered by these centers. For the latter type of data the Component Metadata Infrastructure (CMDI) framework (www.clarin.eu/cmdi/, (Broeder, et al., 2010)) was developed. In CMDI, ISOcat is a key provider of semantic information and forms the basis for semantic interoperability of a wide diversity of metadata profiles. Potentially ISOcat could play the same role for semantic interoperability of language resources, i.e., allowing explicit semantics of the data. This can help to provide more advanced ways to find data for both humans and machines, i.e., by overcoming terminological and data organization differences.

This paper starts with a background section on ISOcat and continues to describe the main experiences obtained in many curation and demonstration projects in CLARIN-NL/VL that have used ISOcat to make the semantics of both metadata profiles and language data explicit.

2. ISOcat background

Data categories as used by TC 37 are based on data elements as defined by the ISO 11179 standards (ISO 11179-1, 2004). In the framework of this family of standards a data category is basically a concept with additional specification of its representation, i.e., does the data category have a value domain (complex data category) or not (container or simple data category), and if so what kind of domain (open, closed or constrained) and of which data type.

The ISOcat DCR is the result of an ongoing effort by TC 37 to standardize data categories (Kemps-Snijders, Windhouwer, Wittenburg, & Wright, 2009). Originally these efforts were targeted at the terminology community, which has a long-term tradition in using data categories in the design and exchange of term bases (Wright, 2001). As a successor of the paper list of data categories in ISO 12620:1999 (ISO 12620, 1999) and all the shortcomings of that, i.e., hard to extend with new data categories needed by the community, it was decided to create a registry. The data model of and procedures around the registry are described in ISO 12620:2009 (ISO 12620, 2009).

Although ISOcat is a completely new implementation of this standard, it is also the successor of SYNTAX a pilot DCR implementation, which was based on an early draft (Kemps-Snijders, Ducret, Romary, & Wittenburg, 2006). The data categories stored in SYNTAX and also the registered user base were transformed and imported into ISOcat.

On the one hand this underlines the intention and obligation of TC 37 to keep the past data category specification work available. But on the other hand the quality of a lot

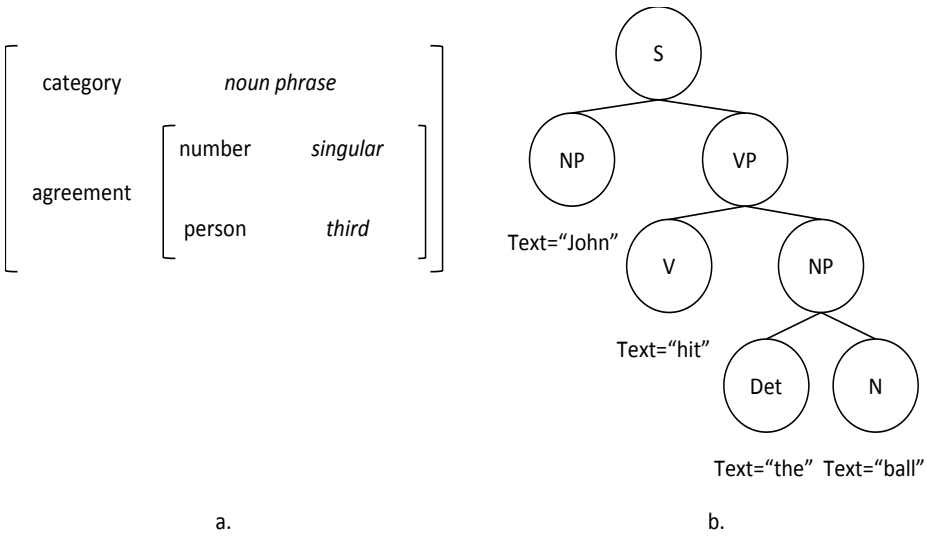

Figure 1: The use of the noun phrase concept a) as a simple data category in a feature structure and b) as a container data category in a parse tree. (Windhouwer, 2012)

of these inherited specifications is also considered problematic. Uptake of ISOcat by new users is hampered when they inspect such sub-optimal entries, either inherited or recent additions.

The ISO 12620:2009 standard and its implementation in ISOcat have been largely driven by the requirements of ISO TC 37 and far less so by CLARIN, as design and development of the CLARIN infrastructure was just starting up. However, the CLARIN infrastructure has been growing the last few years and its own requirements have now become clearer. This process has been partially due to the experiences with ISOcat described in this paper.

3. ISOcat experiences

In this section several topics are discussed, which have led over time to a shift of focus for and the usage of the DCR.

3.1 Steady growth of use

Since its launch early 2008 the user base has grown from around 100 to over 500[1]. Also the number of data categories has grown: from just below 2,000 to more than 5,000. These data categories are owned by just a quarter of the user population, where the average is the stewardship for 45 data categories. The last year the average number of requests for data category specifications was 550 a day (see (Wright, Windhouwer, Schuurman, & Kemps-Snijders, 2013) for more statistics).

3.2 Standardization

The previous section gave statistics, which show data category specification activity of individual users in the registry. In the original design these specification activities would ultimately feed the standardization activities of TC 37, i.e., in the form of peer review by Thematic Domain Groups (TDGs). Unfortunately, although some TDGs, i.e., the ones for Metadata, Morphology and Terminology, have been quite active, this did not result in any standardized data categories yet.

One would expect that new standards produced by ISO TC 37 would help to drive the need for standardized data categories within the community. But uptake of ISO 12620:2009, also within non-CLARIN related ISO standardization activity, has been problematic. For example, both the Lexical Markup Framework (ISO 24613, 2008) and the Linguistic Annotation Framework (ISO 24612, 2012) refer users to ISOcat but fail to clarify to the users of these standards how to actually embed data category references in their models. The recently released the Morpho-sytactic Annotation Framework (ISO 24611, 2012) standard contains an appendix listing data categories taken from ISOcat but, again, without references.

3.3 Community efforts

Because of the lack of standardization of proposed data categories it became more urgent for CLARIN-NL/VL to create possibilities for 'community approved' data categories within ISOcat (Wright, Windhouwer, Schuurman, & Kemps-Snijders, 2013). Existing features in ISOcat, e.g., user groups, were extended for this purpose. A group like CLARIN-NL/VL can now recommend data categories to its user community. To accommodate a cleaned up view on the registry, avoiding the content of other efforts, ISOcat can now be started in a group specific mode where only data categories selected by group members are shown. To guide these community efforts a coordinator [2] was appointed, who gives guidelines and reviews the data categories before recommending them on behalf of the community. The CLARIN-NL/VL group can thus provide its community with a clear entry point into the registry.

3.4 Suitability of the data model

In the model, the representation part of a data category specification causes several kinds of problems, which will

[1] Statistics collected in September 2013.

[2] Ineke Schuurman, who is a co-author of this paper.

be discussed in this section.

Proliferation due to types: Although the concept is the same if the representation differs (see Figure 1 for an example), another data category has to be created, i.e., the model does not cater for sharing the same concept across data categories with different representations (cf. above), which leads to perceived proliferation in the registry.

Conflicts with regard to the representation: Data categories are embedded in a resource context, i.e., in the resource schema or in the resource instance itself, and in this context the representation information is already available making the information in the data category specification redundant and possibly conflicting. In the joint CLARIN metadata domain several of these conflicts can be found. In principal there is an intuitive correspondence between the various building blocks of CMDI metadata profiles, i.e., components match with container data categories, elements match with complex data categories and values with simple data categories. However, recent statistics[3] of the mapping show conflicts:

- 165 elements and 72 components are linked to simple data categories;
- 778 components are linked to complex data categories;
- 4 elements are linked to container data categories.

This indicates that metadata modellers ignored the representation information from the data category specification and just selected them on basis of their matching semantics.

Demands a rare blend of expertise: Although some persons do combine linguistic and technical expertise, in many projects these are different roles by different people, but to create a correct data category specification both aspects need to be addressed and this can be problematic.

A core set of ISO standardized data categories ready to be used by the CLARIN community might have made the complexity of the data model acceptable, but now the complexity is perceived as a stumble block to achieve the desired level of semantic interoperability. In the next section strategies to lighten this burden are discussed.

3.5 Maintenance and sustainability

The maintenance of ISOcat was handed to the MPI for Psycholinguistics that also serves as its ISO Registration Authority since the end of 2008. Although they were successful in improving considerably on the old SYNTAX implementation and integrated ISOcat in the CLARIN CMD framework, complaints on the current user interface do persist. These usually concern the speed and complexity of the UI. The latter problem is mainly caused by the DCR's complex model. Also the development load of the standardization workflow support has been quite high leaving less time for usability improvements. In all, solving these problems will remain costing efforts next to the normal maintenance costs. The current problems with the uptake of ISOcat and the perceived discrepancy between costs and the yield in accepted and standardized data categories, ask for an evaluation of its current status by the TC 37 and CLARIN communities.

4. CLARIN requirements

As discussed in section 3.4 there might be conflicts between the data category representation in the registered (and possibly standardized) specification and its use in a resource context. TC 37 and the CLARIN practice might have different viewpoints on this: from TC 37's standardization perspective it is natural to see the data category representation as prescriptive and thus that kind of usage as erroneous, while in CLARIN the representation is considered a hint and can be overruled by the local context.

Also, currently, recommendation is often hampered by offenses with regard to the representation information, for example data categories of type simple not being related to a closed value domain, cf. section 3.4.

By leaving the representation information to be made explicit by the context in which it appears, the data category specification can solely focus on its semantic description as can be provided by a linguist. But is it then still a specification for a data category or has it become a concept? The important question is if CLARIN actually needs a concept registry instead of a data category registry? When considering also the desire to try to share such registries with other communities from cost aspects, it looks advantageous to generalise the requirements as much as possible and adopt a semantic registry with a simpler data model.

There also is a difference in view with respect to the requirements to be met by the definitions between the TC 37 community and the CLARIN community: whereas the first wants their definitions to be applicable to as many languages as possible (broad definitions), the latter need definitions that describe their use in a more specific context (explicit definitions). Note that this will remain a cause for proliferation of the number of entries, even in a simpler semantic registry.

However any type of registry used, should also have sufficient support for community coordination. It has become clear that community efforts like CLARIN require a finer grained and faster system of agreements with possibly limited scope, feedback and coordination than the ISO standardisation procedure can offer.[4] And although ISOcat has been adapted somewhat to support this, the real efforts will need to come from human coordination, e.g., by multiple national CLARIN ISOcat content coordinators.

5. Data Concept Registry

At the end of 2014 the CLARIN and TC 37 communities have met to discuss the future of ISOcat as a Data Category Registry (www.isocat.org/2013-SR/). In this con-

[3] Statistics from December 2013. Thanks to Matej Durco.

[4] Besides, not only ISO standards are included in ISOcat. Some standards have other backgrounds (like EAGLES), or are not full standards, but act as *de facto* or pseudo-standards, for example within a specific language.

structive meeting there was an acknowledgement that the ISO 12620:2009 data model and standardization procedure did not meet the expectations of the broad intended audience. Especially within the language research data community continuation of the current system may hamper further work on promoting semantic interoperability using semantic registries. This also because the ISOcat approach is difficult to share with other research communities, which would be advantageous with regard to cost sharing.

Therefore it was agreed that CLARIN will investigate alternative solutions for a semantic registry with a simplified data model and simplified review and vetting procedures. The simpler data model will focus on describing the semantics of a concept and the location of representation information is left to the communities using the concept. Hence the new registry is named a Data Concept Registry. Although the focus is thus on concepts the new registry model can potentially offer room to store additional information, e.g., the representation information using a generic extension mechanism. Such additions would then be a choice of a particular concept modeller and not be enforced by the model or the registry itself.

The Data Concept Registry will not provide full support for the ISO standardization workflow. Instead communities can recommend concepts, as already is possible in ISOcat for data categories. If wanted, and TC 37 chooses to adopt the new registry as an ISOcat replacement, it may use the same facilities to mark concepts as officially recommended by them.

Next to working out the requirements of the new registry in more detail, an important issue is to investigate the possibility to implement the new registry with off-the-shelf software. An important lesson of the ISOcat experience has been that it is more advantageous to fund content management rather than software development and the hope is that using existing software will limit the implementation and maintenance load.

6. Conclusion

In all the work on ISOcat with ISO TC37 has brought inspiration how semantic interoperability issues could be handled with light weight ontology type of means, and what costs and efforts are involved. However the realization of the problematic match of ISOcat with regard to the more dynamic research community workflow requirements was slow to come. More frequent evaluation of the uptake might have brought the above explained decision earlier.

The experience with using ISOcat and ISO defined standardization workflow within the CLARIN project, that is primarily directed to research data has shown us that it is needed to bring more distance between the dynamics of the research community and the necessarily more static and slow flow of the ISO work.

References

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., et al. (2010). A Data Category Registry- and Component-based Metadata Framework. *Seventh International Conference on Language Resources and Evaluation.* Malta: ELRA.

ISO 11179-1. (2004). *Information technology -- Metadata registries (MDR) -- Part 1: Framework.* Geneve: International Organization for Standardization.

ISO 12620. (1999). *Data Categories.* Geneve: International Organization for Standardization.

ISO 12620. (2009). *Specification of data categories and management of a Data Category Registry for language resources.* Geneve: International Organization for Standardization.

Kemps-Snijders, M., Ducret, J., Romary, L., & Wittenburg, P. (2006). An API forAccessing the Data Category Registry. *Fifth International Conference on Language Resources and Evaluation.* ELRA.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. (2009). ISOcat: Remodeling Metadata for Language Resources. *International Journal of Metadata, Semantics and Ontologies*, 261-276.

Windhouwer, M. (2012). RELcat: a Relation Registry for ISOcat data categories. *Eight International Conference on Language Resources and Evaluation.* Istanbul, Turkey: ELRA.

Wright, S. E. (2001). Data Categories for Terminology Management. In *The Handbook of Terminology Managemen* (pp. 552-571). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Wright, S. E., Windhouwer, M., Schuurman, I., & Kemps-Snijders, M. (2013). Community efforts around the ISOcat Data Category Registry. In I. Gurevych, & J. Kim, *The People's Web Meets NLP: Collaboratively Constructed Language Resources* (pp. 349-373). Berlin, Heidelberh: Springer.