# RELcat: a Relation Registry for ISOcat data categories

**Menzo Windhouwer**
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
Menzo.Windhouwer@mpi.nl

**Abstract**

The ISOcat Data Category Registry contains basically a flat and easily extensible list of data category specifications. To foster reuse and standardization only very shallow relationships among data categories are stored in the registry. However, to assist crosswalks, possibly based on personal views, between various (application) domains and to overcome possible proliferation of data categories more types of ontological relationships need to be specified. RELcat is a first prototype of a Relation Registry, which allows storing arbitrary relationships. These relationships can reflect the personal view of one linguist or a larger community. The basis of the registry is a relation type taxonomy that can easily be extended. This allows on one hand to load existing sets of relations specified in, for example, an OWL (2) ontology or SKOS taxonomy. And on the other hand allows algorithms that query the registry to traverse the stored semantic network to remain ignorant of the original source vocabulary. This paper describes first experiences with RELcat and explains some initial design decisions.

**Keywords:** linguistic data categories, ontological relationships, relation registry

## 1. Introduction

Since early 2009 the ISOcat Data Category Registry (DCR) is operational[1]. A DCR is an ISO 12620:2009 compliant registry for elaborate specifications of data categories (ISO 12620, 2009). Data categories are elementary descriptors in a linguistic resource or annotation scheme. The predecessor of ISOcat SYNTAX, which was based on a draft of the revised ISO 12620 DCR data model allowed some relationships to be included in the specification (Ide and Romary, 2004): 1) *value of* relationships between simple and closed data categories, and 2) *broader generic concept* relationships between data categories. The final DCR data model restricted the *broader generic concept* relationship to be between simple data categories only (Kemps-Snijders, Windhouwer et al., 2008). The reasons for these were:

1. It would already be hard to agree on specifications of standardized data categories, let alone of putting them into a single ideal linguistic ontology,
2. Data categories would be put in ontological relationships based on very domain or user specific views thus hampering reuse in different contexts, and
3. Allowing multiple ontologies to coexist in the registry could lead to endless ontological clutter.

These are valid reasons and highlight the goal of the DCR to come to a standardized and reusable set of data categories. However, it was already clear that in the end ontological relationships, especially based on equivalence, would be important to provide crosswalks between various (application) domains and/or between various registries. Now that ISOcat gets nearer to a stable release with relatively complete functionality [2] a companion registry is under development whose aim is to store all kinds of relationships and manage the ontological clutter. This Relation Registry is called RELcat (Schuurman and Windhouwer, 2011). This paper will report on the first experiments with RELcat and explain some initial design choices. But the next section will first stress the need for this new registry based on experience gained with the use of ISOcat.

## 2. Proliferation of Data Categories

In an ideal world a linguistic concept would be represented by a single data category, whose specification can then be standardized and used by the linguistic community as a whole. However, this idealistic view is already compromised by the DCR data model itself. In this data model data categories can be of various types:

- *Complex data categories* which have a conceptual domain,
- *Simple data categories* which are values in such a conceptual domain, and
- *Container data categories*[3] which group other container or complex data categories.

Which data category type is appropriate is directly related to how the data category is used in a linguistic resource. For example, the linguistic concept *noun phrase* can be realized as either a container data category or a simple data category. This is illustrated in Figure 1[4] where in the feature structure *noun phrase* is a value of the *category* attribute, i.e., a simple data category which is part of the conceptual domain of a complex data category, and in the parse tree *NP* (a common abbreviation for *noun phrase*) is an inner node, i.e., a container data category.

[1] See http://www.isocat.org/

[2] Active usage of ISOcat has of course revealed shortcomings which will have to be dealt with in subsequent versions, but version 1.0 will be functionally complete with regard to the original aims.

[3] The container data category type isn't part of the data model in ISO 12620:2009, but its addition to ISOcat has been sanctioned by ISO TC 37.

[4] Source of the feature structure: http://en.wikipedia.org/wiki/Feature_structure, and source of the parse tree: http://en.wikipedia.org/wiki/Bottom-up_parsing

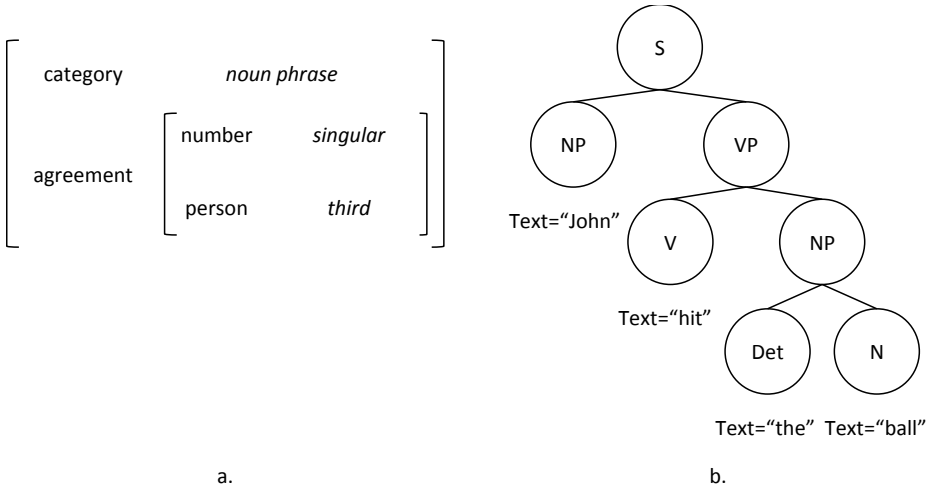


Figure 1 The use of the noun phrase concept in a) a feature structure and b) a parse tree

The fact that these two data categories which both realize the concept noun phrase are closely related is valuable information which can be exploited by, for example, semantic search algorithms to find closely related resources. The Relation Registry, RELcat, allows storing and typing these kinds of relationships.

Another source of proliferation is the fact that existing sets are bulk loaded into ISOcat, e.g., the NKJP (Patejuk and Przepiórkowski, 2010) or STTS tagsets or the concepts of the GOLD ontology (Farrar and Langendoen, 2010). Ideally these existing sets would all reuse the same concepts and indicate their own encoding using Data Element Names, e.g., for the data category /*adverb*/ (DC-1232) the following Data Element Names could exist:

1. Name *adverbial*, source: GOLD
2. Name *adv*, source: NKJP
3. Name *ADV*, source STTS

However, there is currently no data category that is shared this way. Either due to fundamental differences between data categories, which seem equivalent just based on their names but are deemed deviant enough by the linguist mapping these sets to already existing ISOcat data categories. There might be a stronger drive for harmonization if there was already a standardized core of data categories in ISOcat. However, at the moment all data categories are owned by users and none are standardized and owned by a Thematic Domain Group. Also there is the perceived need to own all data categories which are related to a specific tagset or resource type, and this is indeed a valid concern if the owner of the equivalent data categories is not responding to requests to include the required Data Element Name.

The ultimate source of proliferation of data categories in ISOcat is in the end that the registry is based on a grass roots approach, i.e., each user can add data categories that he or she needs. The idea is that in due time, when these data categories have matured, they can be submitted for standardization to one of the Thematic Domain Groups and a coherent, but extensible, core of standardized data categories will appear. However, due to this open approach it is also fairly easy for users to create data categories which are (almost) equivalent to already existing data categories.

## 3. RELcat: a Relation Registry

Originally envisioned to be a registry to allow crosswalks between data categories from various registries the role of the Relation Registry has become more important due to the proliferation of data categories within one registry as sketched above. Also it is clear that various (competing) ontological relationships between data categories need to be stored. Driven by the increasing need of storing these relationships the development of a Relation Registry, named RELcat, has started. The natural form of these relationships is a triple of:

1. *Subject*: the data category the statement is about;
2. *Predicate*: the relationship between the subject data category and the object data category;
3. *Object*: the data category the relationship is directed at.

As there is the possibility of competing ontological relationships, i.e., private views of specific users, the user should be able to indicate which sets of relationships she wants use. This accommodated by extending the triple to a quad:

4. *Set*: the set of which this relationship statement is a part.

These requirements meet the specifications of a RDF quad store. The current quad store RELcat uses is OpenAnzo[5] which supports queries that indicate which graphs are to be considered.

The subject and object URIs in the Relation Registry are obvious: these are the Persistent IDentifiers of the ISOcat data categories or general URLs for concepts from other registries (Windhouwer and Wright, 2012). However, which predicates to use? OWL (W3C, 2009) and SKOS (W3C, 2009) are prime candidates to provide the predicates. But linguistic knowledge resources are already available in various formats, which might not all be readily translated to these formal defined predicates. Also these formal predicates are defined between specific subclasses of RDF resource, e.g., RDF class or RDF property. For example in OWL the predicate *owl:equivalentProperty* can only be used between to RDF resources of type RDF property. As one of the aims is that

[5] See http://www.openanzo.org/

RELcat doesn't prescribe design choices for linguistic knowledge bases the taken approach tries to be maximally flexible. It does so by building on the following core taxonomy of relationship types:

1. Related (*rel:related*)
1.1. Same as (*rel:sameAs*)
1.2. Almost same as (*rel:almostSameAs*)
1.3. Broader than (*rel:broaderThan*)
1.3.1. Super class of (*rel:superClassOf*)
1.3.2. Has part (*rel:hasPart*)
1.3.2.1. Has direct part (*rel:hasDirectPart*)
1.4. Narrower than (*rel:narrowerThan*)
1.4.1. Sub class of (*rel:subClassOf*)
1.4.2. Part of (*rel:partOf*)
1.4.2.1. Direct part of (*rel:directPartOf*)

Now multiple vocabularies for predicates can be supported by adding them to their proper place in this taxonomy. For example, the diverse OWL and SKOS equivalence predicates can be supported as follows:

1.1. Same as (*rel:sameAs*)
1.1.1. OWL same as (*owl:sameAs*)
1.1.2. OWL equivalent class (*owl:equivalentClass*)
1.1.3. OWL equivalent property (*owl:equivalentProperty*)
1.1.4 SKOS exact match (*skos:exactMatch*)
1.2 Almost same as (*rel:almostSameAs*)
1.2.1 SKOS close match (*skos:closeMatch*)

The same can be done for the other OWL and SKOS predicates, as well as any other existing vocabulary. The benefit of this approach is that

1. existing linguistic knowledge bases can be loaded into or accessed by RELcat as they are, and
2. generic algorithms can use the relationships without intimate knowledge of all these different vocabularies.

Relationships between ISOcat data categories and Dublin Core metadata elements can be, for example, specified as follows:

```
@prefix relcat : <http://www.isocat.org/relcat/set/> .
@prefix rel : <http://www.isocat.org/relcat/relations#> .
@prefix dc : <http://purl.org/dc/elements/1.1/> .
@prefix isocat : <http://www.isocat.org/datcat/> .

relcat:cmdi {
    isocat:DC-2573 rel:sameAs dc:identifier .
    isocat:DC-2482 rel:sameAs dc:language .
    ...
    isocat:DC-2556 rel:subClassOf dc:contributor .
    isocat:DC-2502 rel:subClassOf dc:coverage .
}
```

A semantic search algorithm that broadens a search for linguistic resources by looking for fuzzy equivalence can express its queries completely in the RELcat vocabulary using, for example, *rel:sameAs* and *rel:almostSameAs*.

For example, the following SPARQL query will find equivalence relationships involving the */languageID/* (DC-2482) data category:

```
PREFIX rel: <http://www.isocat.org/relcat/relations#>
PREFIX cat: <http://www.isocat.org/datcat/>

SELECT ?rel WHERE { cat:DC-2482 rel:sameAs ?rel . }
```

The result will include the Dublin Core metadata elements, but may, when the GOLD relation set is selected, also include GOLD concepts. where relationships are defined in an OWL ontology annotated with ISOcat data category references. But it also still possible to access the original predicates and, for example, combine OWL-based relation sets and feed it into an OWL reasoner.

The current alpha version of RELcat supports basic storage and retrieval of relationship sets based on this taxonomy in the OpenAnzo quad store. It also supports SPARQL query templates (so the actual SPARQL is hidden from the user) on (combinations of) relation sets. Sets currently available in RELcat include:

1. relationships between ISOcat metadata categories and Dublin Core metadata elements as created for CMDI (Broeder, Uytvanck et al., 2012), and
2. relationships between ISOcat categories and GOLD concepts as created for the RELISH project (Aristar-Dry, Drude et al., 2012).

Although the current version is missing a (web) user interface to add new relationships the collection of sets is already expanding, which indicates the increasing the demand for this kind of registry.

## 4. Future work

The current development focuses on supporting and exploiting properties of relation types, like symmetric, transitive and inverses, which will allow the user to specify only a minimal set of relationships. For this RELcat will need a basic reasoner, e.g., based on RDFS Plus (Allemang and Handler, 2008).

Depending on the future size of ISOcat and RELcat the use of (almost) same as relationships might require special handling. These (loose) equivalence relationships might possibly lead to a combinatorial explosion. Some commercial triple stores already offer specific support to handle these kinds of large graphs:

1. Oracle Database Semantic Technologies (Oracle, 2005) support so called *owl:sameAs* cliques which can be consolidated by choosing a clique representative, and
2. OWLIM (ontotext, 2011) supports a similar approach using a so called master node.

In RELcat the same kind of approach could be followed: 1) either select one representative from the same-as clique or 2) give the same-as clique its own identifier. A possible complication there is the dynamic combination of relationship sets, which might require that these cliques need to be computed on the fly.

## 5. Conclusion

With the increasing usage of ISOcat various causes for proliferation of data categories have appeared. Some of these have fundamental reasons, e.g., due to basic properties of the DCR data model like data category types, and cannot be remedied. To retain semantic interoperability the ontological relationships between these data categories need to be stored in a semantic network to store semantic crosswalks. RELcat is starting to provide the basic facilities to do so, and is specifically aiming at easily integration of already existing linguistic knowledge bases.

## 6. Acknowledgements

Thanks to early adaptors Matej Durco (SMC4LRT), Irina Nevskaya (RELISH) and Ineke Schuurman (CLARIN-NL/VL) for driving this first version of RELcat forward.

## 7. References

Allemang, D. and J. Handler (2008). Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, Morgan Kaufmann.

Aristar-Dry, H., S. Drude, J. Gippert, I. Nevskaya and M. Windhouwer (2012). „Rendering Endangered Lexicons Interoperable through Standards Harmonization": the RELISH Project. Language Resources and Evaluation. Istanbul, ELRA.

Broeder, D., D. v. Uytvanck, M. Windhouwer, M. Gavrilidou and T. Trippel (2012). Standardizing a Component Metadata Infrastructure. Language Resources and Evaluation. Istanbul, ELRA.

Farrar, S. and D. T. Langendoen (2010). An OWL-DL implementation of GOLD: An ontology for the Semantic Web. Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology. A. W. Witt and D. Metzing. Dordrecht, Springer.

Ide, N. and L. Romary (2004). A Registry of Standard Data Categories for Linguistic Annotation. Language Resources and Evaluation. Lisbon, Portugal.

ISO 12620 (2009). Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources, International Organization for Standardization.

Kemps-Snijders, M., M. Windhouwer, P. Wittenburg and S. E. Wright (2008). A Revised Data Model for the ISO Data Category Registry. Terminology and Knowledge Engineering B. N. Madsen and H. E. Thomsen. Copenhagen, Denmark.

ontotext. (2011). owl-sameAs-optimization. from http://www.ontotext.com/owlim/owl-sameas-optimisation.

Oracle. (2005). Oracle Database Semantic Technologies Developer's Guide. 2011, from http://download.oracle.com/docs/cd/E14072_01/appdev.112/e11828/toc.htm.

Patejuk, A. and A. Przepiórkowski (2010). ISOcat Definition of the National Corpus of Polish Tagset. Language Resource and Language Technology Standards workshop at LREC 2010, Malta.

Schuurman, I. and M. Windhouwer (2011). Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? 2nd Supporting Digital Humanities conference. Copenhagen, Denmark.

W3C. (2009). OWL Web Ontology Language. from http://www.w3.org/standards/techs/owl.

W3C (2009). SKOS Simple Knowledge Organization System.

Windhouwer, M. and S. E. Wright (2012). Linking to linguistic data categories in ISOcat. Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata. C. Chiarcos, S. Nordhoff and S. Hellmann, Springer-Verlag: 99-107.