Linguistic resources and cats: how to use ISOcat, RELcat and SCHEMAcat

Menzo Windhouwer[a], **Ineke Schuurman**[b]
[a] The Language Archive – DANS
Anna van Saksenlaan 51, The Hague, The Netherlands
[b] Katholieke Universiteit Leuven, University of Utrecht
Blijde Inkomststraat 13, Leuven, Belgium
Menzo.Windhouwer@dans.knaw.nl, ineke@ccl.kuleuven.be

Abstract

Within the European CLARIN infrastructure ISOcat is used to enable both humans and computer programs to find specific resources even when they use different terminology or data structures. In order to do so, it should be clear which concepts are used in these resources, both at the level of metadata for the resource as well as its content, and what is meant by them. The concepts can be specified in ISOcat. SCHEMAcat enables us to relate the concepts used by a resource, while RELcat enables to type these relationships and add relationships beyond resource boundaries. This way these three registries together allow us (and the programs) to find what we are looking for.

Keywords: registries, sharing explicit semantics, semantic network

1. Introduction

Ideally, metadata, i.e. data about the data contained in a document, a corpus, etc. should always be available for documents used for research, be it textual, audio or visual media, or a mixture of these. There can be many sorts of metadata associated to a document, e.g., of a bibliographical type (creator, location, publication date, etc.) to more descriptive metadata (recording date, age of interviewees, place of residence, type of annotation, etc.).
What is not so common is to describe the terminology and underlying concepts used in the content of the linguistic resources. There is no fixed set of metadata, i.e. a single metadata schema, to describe all types of resources. And the same holds for the terminology and concepts used. Different communities of practice use different terminology and conceptual frameworks. As a consequence, several terms and concepts are sometimes used to describe the same state of affairs, or, the other way round, the same term or concept is used to describe different states of affairs. So, how will a user know what is meant? Somewhere applicable meanings of specific concepts are to be made clear. Both for the benefit of human users, like a researcher enriching a resource with either metadata or (linguistic) annotations or using data from such an (annotated) resource for research purposes, and for computer programs. If some description of terms and concepts is available it is in general in the form of documentation meant to be read by a human.
When the fundaments of the European CLARIN[1] infrastructure were designed in its preparatory phase it became clear that the semantics of terms and concepts used in both the metadata for and in the content of language resources should be made explicit. This way the infrastructure creates the potential to provide flexible mechanisms to deal with the wide variety to be found in the language resources it would deal with. Around the same time ISO Technical Committee 37 *Terminology and other language and content resources* (ISO TC 37) started on a reimplementation of its Data Category Registry (DCR), named ISOcat. As promoting the use of standards is one of the aims of CLARIN and the DCR could function as one of the corner stones for semantic interoperability CLARIN supported the development of ISOcat and promoted its use by its community.
In this paper we will pay attention to some content, i.e. not metadata, related issues arising in ISOcat, mention some of the adaptations the way we deal with them and introduce some additions: the companion registries RELcat and SCHEMAcat. Using the CGN part of speech tagset (Van Eynde, 2004) as a running example, we will show that all these registries together allow a complete semantic description of this resource. It will also show that many of the basic building blocks can be shared by the semantic description of other resources, that crosswalks can be created and thus semantic interoperability is fostered.

2. ISOcat, the Data Category Registry

ISOcat is an ISO 12620:2009 compliant registry (ISO 12620, 2009) in which such concepts, in the context of this registry called data categories (see also (Broeder, Schuurman, & Windhouwer, 2014)), and various terms used for them are described in a concise way. In general the descriptions are meant to be useful for as many users as possible; on the other hand specific uses might require very specific readings of a concept. This might lead to the creation of multiple data categories, which are still semantically close. It is relatively easy to create new data categories, as ISOcat is rather liberal and open: *In se* everybody can submit entries, and very few rules are to be obeyed when doing so. However, it turns out that some of these few rules presented problems for uninitiated users (see once more Broeder, Schuurman & Windhouwer (2014)). Deployment of ISOcat within the CLARIN

[1] Cf. http://www.clarin.eu.

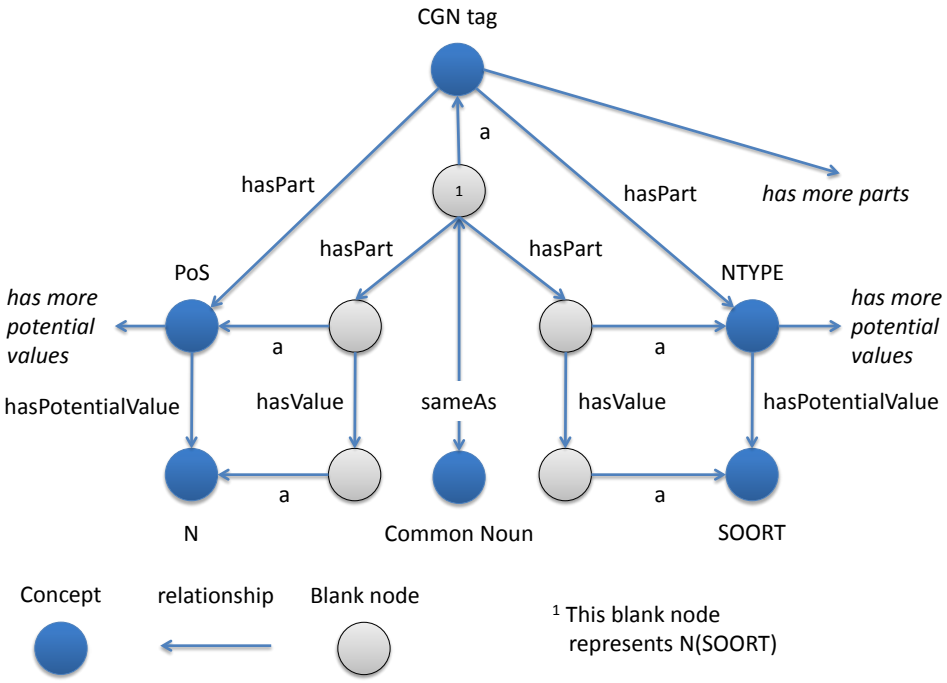

Figure 1: Mapping N(SOORT) to /*Common Noun*/

infrastructure means that the CLARIN community[2] imposes some additional rules and creates a more controlled and coherent CLARIN subset within the registry. Some of these rules are:

- The definitions used should be unambiguous. i.e., when in a definition a specific linguistic concept is used, this concept is to refer to its definition elsewhere in ISOcat, by mentioning its persistent identifier, preferably in the Note Section coming with the definition (see (Kemps-Snijders, Windhouwer, Wittenburg, & Wright, 2008) for a more detailed description of the DCR data model);
- Definitions should not be circular;
- Do not create a new data category when another one can be adopted;
- Take care that the definition in the working language (English) and other languages still match after small improvements have been made;
- A definition of a public data category cannot be changed in a meaningful way: if necessary a new entry is to be created, the old one being made superseded/deprecated;
- Do not make a data category private, once it has been public;
- Definitions should be to the point (one or two sentences);
- Definitions are to be as language and theory independent as possible in order to promote their reusability;
- Specify relevant abbreviations and other more technical identifiers for a data category in the Data Element Name section;
- Use the proper English spelling, and use camelCase whenever asked for.

For the CGN tagset data categories have been created[3]. These data categories describe the basic building blocks of these tags, i.e., the parts of speech and features and feature values that describe them. Care has been taken to define them in a way that makes the data categories reusable, i.e., specific CGN aspects like the labels (in many cases abbreviations) are specified in the Data Element Name sections but the definition is kept generic. This enables other tagsets or different kinds of linguistic constructs to reuse these data categories.

In the end, we want to avoid, as far as possible, proliferation of data categories. But still sometimes it is needed that near equivalent data categories are added.[4] In the Section 3 it is shown how RELcat can be of help in order to relate such data categories. But first the next section will describe how data categories can be linked to their instantiations in a resource.

(Broeder, Schuurman, & Windhouwer, 2014) describes the major changes that are envisioned for a next version of ISOcat, where the focus will shift from data categories to concepts. The impact of these changes will be limited with regard to the interaction between the various registries described in this paper.

3. Publishing a schema in SCHEMAcat

Data categories are defined as "the result of the specification of a given data field" (ISO 12620, 2009). Instantiations of these data fields appear in linguistic resources. Although it is possible to annotate individual resources with references to the data categories they use, it is more efficient to annotate the schema of these resources, i.e., this way a large number of resources can be annotated in one go (Windhouwer & Wright, 2010).

[2] The Dutch and Flemish national CLARIN initiatives (also known as CLARIN-NL/VL) are doing some pioneering work in this respect.

[3] See http://www.isocat.org/rest/dcs/530

[4] Causes for proliferation are mentioned in (Windhouwer, 2012).

SCHEMAcat is a schema registry[5] which will become a place to persistently store schemas of various kinds, e.g., XML schemas for XML-based resources but also text-oriented schemas like EBNF grammars. The registry will support long term storage of schemas, i.e., decoupling a schema from a relatively temporary tool, service or project and safeguarding its existence for archived resources based on this schema. In the workflow to make a schema available in SCHEMAcat a persistent identifier, i.e., a handle, is assigned to a schema (ISO 24619, 2011). This handle needs to be included in the resources, if applicable, or at least their metadata.

As described above, ISOcat contains data categories that describe the basic building blocks of a CGN tag. How these blocks are combined in a complete tag can be described in an EBNF grammar, which then functions as a schema for CGN tags. An example snippet of this grammar is shown in Figure 2. This example also illustrates the embedding of the references to the ISOcat data categories in the schema.

```
(* @dcr:datcat 'WW'  isocat:DC-4949 *)
(* @dcr:datcat 'TW'  isocat:DC-4950 *)
tag    = WW '(' WVORM ',' … ')'
       | TW … ;
…
(* @dcr:datcat WVORM isocat:DC-4957 *)
(* @dcr:datcat 'buigbaar' isocat:DC-4960 *)
(* @dcr:datcat 'infinitief' isocat:DC-4961 *)
WVORM = 'buigbaar' |
         'infinitief' | … ;
…
```

Figure 2: CGN schema snippet[6]

This EBNF schema received a handle, i.e., hdl:1839/00-SCHM-0000-0000-004A-A, from SCHEMAcat.[7] The schema registry also stores metadata about the schema in which, for example, specific characteristics of a tagset can be described.

In the context of exploiting the semantic network created by the combination of ISOcat, RELcat and SCHEMAcat for semantic searches the function of SCHEMAcat will be on two levels:

1. identify candidate matching archived resources based on their SCHEMAcat schema containing patterns of specific relationships between specific data categories, and
2. check that the resource actually instantiates this pattern by running a specific validation method.

The second level might not be possible for all schema types, but SCHEMAcat should allow to plugin these specific validation methods. For example, an EBNF plugin would be able to parse the CGN tag `WW(persoonsvorm, verleden, enkelvoud)` and return the set of used data categories, i.e., {*/verb/* (isocat:DC-4949), */partOfSpeech/* (isocat:DC-1345), */finite/* (isocat:DC-4958), */finiteness/* (isocat:DC-1893), */past/* (isocat:DC-4966), */tense/* (isocat:DC-4964), */singular/* (isocat:DC-4974), */number/* (isocat:DC-4916)}, which is just a small subset of all the data categories associated with the CGN tagset. Also only some of these data categories and the ontological relationships they have among each other may be specified in RELcat.

4. Relating data categories in RELcat

Sometimes specific needs might lead to the creation of different data categories, which can still be considered equivalent or at least almost equivalent from a semantic viewpoint. This and other ontological relationships between data categories are very valuable as they provide means to, for example, enable a semantic search algorithm to broaden its scope and also return close matches.

Early in the design of ISOcat it was decided that ontological relationships cannot be a part of the standardized core of the registry (Kemps-Snijders, Windhouwer, Wittenburg, & Wright, 2008), i.e., these kinds of relationships are in many cases too specific for an application or resource and might thus hinder reusability.[8] A companion registry to ISOcat, a Relation Registry called RELcat[9], will support the storage of (user-specific) sets of relations.

The two snippets in Figure 3 show various relationships for CGN data categories. The first relationship is between two data categories for use-mention distinction, which are considered semantically almost equal. This is a mapping from a CGN data category to a data category based on TEI and can thus function (TEI Consortium, 2014) as a crosswalk from one resource type to another. The second relationship is more internal to CGN. In the snippet in Figure 2 a slight indentation of the "infinitief" value with regard to the "buigbaar" value hinted at a subsumption relationship between the two values. However, in the EBNF grammar this relationship cannot be made explicit, but RELcat does allow this.

```
relcat:cgn {
 isocat:DC-5034 rel:almostSameAs isocat:DC-364.
}

relcat:cgn-isa {
 isocat:DC-4961 rel:subClassOf isocat:DC-4960.
}
```

Figure 3: Snippets of CGN relations

This example also shows that the relations are semantically typed and this allows the registry to do a limited amount of reasoning with them, e.g., the reverse

[5] At the time of writing SCHEMAcat is under construction, a beta version is available at http://www.isocat.org/schemacat-test/.

[6] Due to space limitations the full URLs are abbreviated, i.e., isocat:DC-4949 is actually http://www.isocat.org/datcat/DC-4949.

[7] In ISOcat, itself a DC with PID http://www.isocat.org/datcat/DC-6159 refers to this schema.

[8] Due to legacy reasons the DCR data model does support *is a* typed relationships between simple data categories. This relationship type is equivalent to the *sub class of* type in RELcat. Notice, that a simple data category can only take part in one subsumption hierarchy in ISOcat but in a theoretically unlimited number in RELcat.

[9] At the time of writing RELcat is under construction, a beta version is available at http://www.isocat.org/relcat-test/.

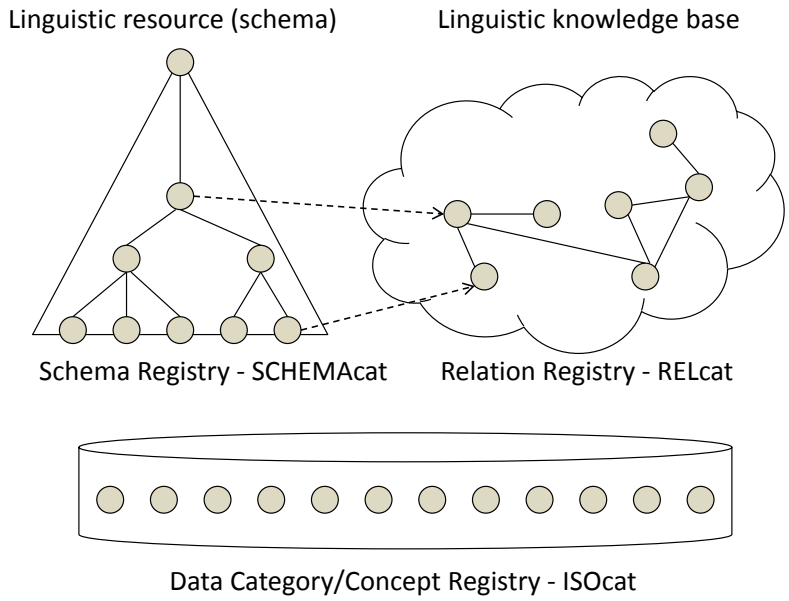

Figure 4: Semantic network created by the cats

of the shown *almost same as* relation is also considered a valid relationship.
Currently the following relationship types are supported:

1. *related*
   - 1.1. *same as* (a symmetric and transitive relationship)
   - 1.2. *almost same as* (a symmetric relationship)
   - 1.3. *broader than* (transitive, and the inverse of the *narrower than* relationship)
     - 1.3.1. *superclass of* (transitive, and the inverse of the *subclass of* relationship)
     - 1.3.2. *has part* (transitive, and the inverse of the *part of* relationship)
   - 1.4. *narrower than* (transitive, and the inverse of the *broader than* relationship)
     - 1.4.1. *subclass of* (transitive, and the inverse of the *superclass of* relationship)
     - 1.4.2. *part of* (transitive, and the inverse of the *has part* relationship)
   - 1.5. *has potential value* (the inverse of the *is potential value of* relationship)
     - 1.5.1. *has value* (the inverse of the *is value of* relationship)
   - 1.6. *is potential value of* (the inverse of the *has potential value* relationship)
     - 1.6.1. *is value of* (the inverse the *has value* relationship)

The value domain relationship types (1.5 and 1.6) are orthogonal to the equivalence and narrower/broader relationship types, as they allow reflecting a specific representation in the relation registry whereas the ontological relationships are not representation specific. Later on it will be shown that these two can interact to specify more complex mappings.
These relationship types are placed in an extensible taxonomy. This means that other relationship types from other vocabularies, e.g., OWL or SKOS, can be inserted at their proper place in this subsumption hierarchy and sets of relations using these vocabularies can be loaded in RELcat. For example, *owl:sameAs*, *owl:equivalentClass* and *owl:equivalentProperty* can be assigned as subtypes of *same as* in the RELcat taxonomy. This allows RELcat to load ontologies and taxonomies expressed in OWL or SKOS and still query them based on the general relation types. For example, the GOLD ontology (Linguist List, 2014) has been imported in RELcat in this way.
From the viewpoint of RELcat the annotated schemas in SCHEMAcat are also a place to harvest relationships between data categories. Depending on the schema type automatic ontological typing of the relationships can be possible. In the other cases only the generic *related* relation type is possible, but users should always be able to select a more specific relationship. So, returning to the CGN example, RELcat can access the CGN EBNF grammar and request the relationships, where at least the value domain relationships can be typed. For an EBNF schema type this boils down to the arcs between nodes in all the possible abstract syntax trees.
More recently work has started on more advanced mappings, e.g., needed due to granularity differences where multiple data categories or concepts play a role on one or both sides of the relationship. The example in Figure 1 shows how a set of CGN tags, i.e., the ones with value N for the PoS slot and with value SOORT for the NTYPE slot can be mapped to a Common Noun data category. The concept nodes all exist in ISOcat[10] and are bound to resources, potentially via a schema, i.e., in this case the CGN EBNF for most of the concepts. The RELcat relations between these nodes reflect their relations among each other in the schema. On the schema level these are represented by *has potential value* relationships. On the instance level, e.g., in the content of a resource one specific value will appear. The instance level is reflected by the blank nodes, i.e., these are

[10] The following data category references have been omitted from Figure 1 to improve its readability */N/* (isocat:DC-4909), */SOORT/* (isocat:DC-4910) */NType/* (isocat:DC-4908), */PoS/* (isocat:DC-5294), */Common Noun/* (isocat:DC-1256), */CGN tag/* (isocat:DC-6159).

anonymous but are typed with a concept. Using this mapping a query for common noun can be extended to CGN tags which fit the mapping to N(SOORT).

5. Conclusion and future work

This paper discussed how for language resources, especially the ones in the CLARIN infrastructure, optimal use can be made of ISOcat, RELcat, and SCHEMAcat. To achieve a higher level of semantic interoperability, i.e., not only based on finding shared data categories, ontological and contextual information will have to be taken into account. RELcat and SCHEMAcat provide the means to harvest and specify this information in the form of relationships and allow (search) algorithms to traverse the semantic graph thus made explicit (see Figure 4). Using this combination of various registries it has become possible to fully describe a tagset, which was basically impossible in ISOcat alone as the experiment for the Polish NKJP tagset indicated (Patejuk & Przepiórkowski, 2010).

The beta versions of RELcat and SCHEMAcat registries have been populated and used in CLARIN already, but development should be completed to make them production ready.

References

Broeder, D., Schuurman, I., & Windhouwer, M. (2014). Experiences with the ISOcat Data Category Registry. *Ninth International Conference on Language Resources and Evaluation (LREC 2014).* Reykjavik, Iceland: ELRA.

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

ISO 24611. (2012). *Morpho-syntactic annotation framework (MAF).* Geneve: International Organization for Standardization.

ISO 24619. (2011). *Persistent identification and sustainable access (PISA).* Geneve: International Organization for Standardization.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2008). A Revised Data Model for the ISO Data Category Registry. *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008).* Copenhagen, Denmark.

Linguist List. (2014). *General Ontology for Linguistic Description*. Opgeroepen op March 14, 2013, van http://linguistics-ontology.org/

Patejuk, A., & Przepiórkowski, A. (2010). ISOcat Definition of the National Corpus of Polish Tagset. *Language Resource and Language Technology Standards.* Malta: ELRA.

TEI Consortium. (2014). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 2.6.0. January 20, 2014.* TEI Consortium.

Van Eynde, F. (2004). *Part of Speech Tagging en Lemmatisering van het CGN Corpus.* Centrum voor Computerlinguïstiek, KU Leuven.

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

Windhouwer, M., & Wright, S. E. (2010). Referencing ISOcat data categories. *LREC 2010 LRT standards workshop.* Malta: European Language Resources Association.