CLARIN Concept Registry: the new semantic registry

Ineke Schuurman
Utrecht University
University of Leuven
ineke@ccl.kuleuven.be

Menzo Windhouwer
Meertens Institute
menzo.windhouwer@meertens.knaw.nl

Oddrun Ohren
National Library of Norway
oddrun.ohren@nb.no

Daniel Zeman
Charles University in Prague
zeman@ufal.mff.cuni.cz

1 Introduction

One of the foundations of the CLARIN Component Metadata Infrastructure (CMDI; Broeder et al. 2012; clarin.eu/cmdi) is a semantic layer (Durco and Windhouwer, 2013) formed by references from CMDI components or elements to entries in various semantic registries. Popular have been references to the metadata terms provided by the Dublin Core Metadata Initiative (DCMI; dublincore.org) and the data categories provided by ISO Technical Committee 37's Data Category Registry (DCR; ISO 12620, 2009), ISOcat (isocat.org). Although using ISOcat has been encouraged by CLARIN, it has its drawbacks. As pointed out by Broeder et al. (2014) and Wright et al. (2014), ISOcat, with its rich data model combined with a very open update strategy has proved too demanding, at least for use in the CLARIN context. Among other things, confusion on how to judge whether a candidate ISOcat entry adequately represents the semantics of some CMDI component or element, has led to proliferation far beyond the real need, resulting in a semantic layer of questionable quality. Therefore, when ISOcat, due to strategic choices made by its Registration Authority, had to be migrated and became, for the time being, static, CLARIN decided to look for other solutions to satisfy the needs of the infrastructure. As a result the Meertens Institute is now hosting and maintaining this new CLARIN semantic registry.

This paper motivates and describes the new semantic registry, the CLARIN Concept Registry (CCR), its model, content and access regime, indicating the differences from ISOcat where appropriate. Proposed management procedures for CCR are also outlined, although not in detail.[1]

2 An OpenSKOS registry

In CLARIN-NL the Meertens Institute had already developed (and continues hosting) the CLAVAS vocabulary service based on the open source OpenSKOS software package (Brugman and Lindeman, 2012; openskos.org), which was originally created in the Dutch CATCHPlus project. The OpenSKOS software provides an API to access, create and share thesauri and/or vocabularies, and also provides a web-based editor for most of these tasks. The software is used by various Dutch cultural heritage institutes. The Meertens Institute joined them to collectively maintain and further develop the software.

Based on the experiences with ISOcat OpenSKOS was evaluated to see if it would meet the needs of the CLARIN community and infrastructure. The major aim was to improve the quality of the concepts by having a much simpler data model and a less open, and also less complicated, procedure for adding new concepts or changing existing ones and recommending them to the community. In addition, certain technological requirements of the CLARIN infrastructure had to be met. Based on this evaluation the Meertens Institute extended the OpenSKOS software in various ways:

- Concepts in the CCR get a handle as their Persistent IDentifier (PID);
- The CCR can easily be accessed by the CLARIN community via a faceted browser;
- Support for SKOS collections;
- Shibboleth-based access to the CCR.

Currently these extensions reside in a private Meertens Institute source code repository, but as part of the CLARIN-PLUS project these extensions (and more) will be integrated with the next version of OpenSKOS now under development.

3 Representing CCR concepts in the SKOS model

The data model supported by OpenSKOS is a substantial part of the Simple Knowledge Organisation Scheme (SKOS) recommendation by W3C (w3.org/skos). SKOS is typically used to represent thesauri, taxonomies and other knowledge organisation systems. At the Meertens Institute support for collections was added and currently Picturae, a Dutch service provider within the cultural heritage world and the original developer of OpenSKOS, works on supporting the extended labels of SKOS-XL.

The work done by the CLARIN community in ISOcat was made available in the CCR by importing selected sets of data categories as new concepts (see Section 4). This made it possible to start a round of clean-up and creating a coherent set of recommended concepts (see Section 5). This import is not lossless as data category specific properties like the data category type and data type are lost. However, these properties have turned out to be one of the main causes of confusion and proliferation in the use of ISOcat (Broeder et al., 2014; Wright et al., 2014). In general SKOS appears to be a suitable model for the CCR. Each CCR concept may be assigned preferred labels (at most one per language), alternative labels, definitions, examples and various kinds of notes. Moreover, the ISOcat thematic domains and data category selections could be maintained by importing them to SKOS concept schemes and collections, respectively. Only one import decision turned out to be problematic: converting the data category identifier into a concept notation. SKOS notations are required to be unique within their concept scheme, whereas this constraint did not apply to data category identifiers in the DCR data model. A first clean-up round to remedy this has been finished successfully.

The SKOS model provides the possibility to express semantic relationships between concepts, e.g. broader than, narrower than and related to. In contrast, the DCR data model did only contain relationships based on the data category types, e.g., a simple data category belonged to the value domain of one or more closed data categories. These domain-range relationships do not correspond well to any of the SKOS relationship types. Careful manual inspection would be needed to determine if any mapping can be made, hence, for now these relationships have not been imported into the CCR. At a later date these facilities of the SKOS model and OpenSKOS can be exploited and could eventually take over the role originally envisioned for RELcat (Windhouwer, 2012). However, for now the initial focus is on the concepts themselves.

Neither SKOS itself nor OpenSKOS does yet provide an extensive versioning model, i.e., concepts can be expired but there is no explicit link to a superseding concept. This is now on the wishlist for the next version of OpenSKOS.

Being RDF-based SKOS also brings the potential to more easily join forces with linked data and semantic web communities.

4 The CCR content

In the past few years, many national CLARIN teams made an effort to enter their data in ISOcat. This work has not been useless as all entries deemed to be worthwhile for a specific CLARIN group were inserted in CCR. Leaving out redundant entries already means a considerable reduction in number of entries (from over 5000 in ISOcat (Broeder et al., 2014) to 3139 in CCR (see Figure 1; clarin.eu/conceptregistry, June 2015)). Although the imported concepts got new handles care was taken to retain a link with their ISOcat origin, so automated mapping is possible and can be used to convert references to ISOcat data categories into references to CCR concepts. A mapping tool for this has been developed and used for the CMDI components, but is generally applicable and available to the CLARIN community (github.com/TheLanguageArchive/ISOcat2CCR).

5 Maintaining the CCR – procedures and actors

Just like ISOcat the CCR can be browsed and searched by anyone, member of the CLARIN community or not, and anyone can refer to the concepts. However, contrary to ISOcat, only specifically appointed users, namely the national CCR content coordinators are given rights to update the CCR. These coordinators were assigned by CLARIN national consortia (see clarin.eu/content/concept-registry-coordinators) when the problems with the usage of ISOcat became apparent, and their mission was to improve the quality of the data categories (now concepts) used

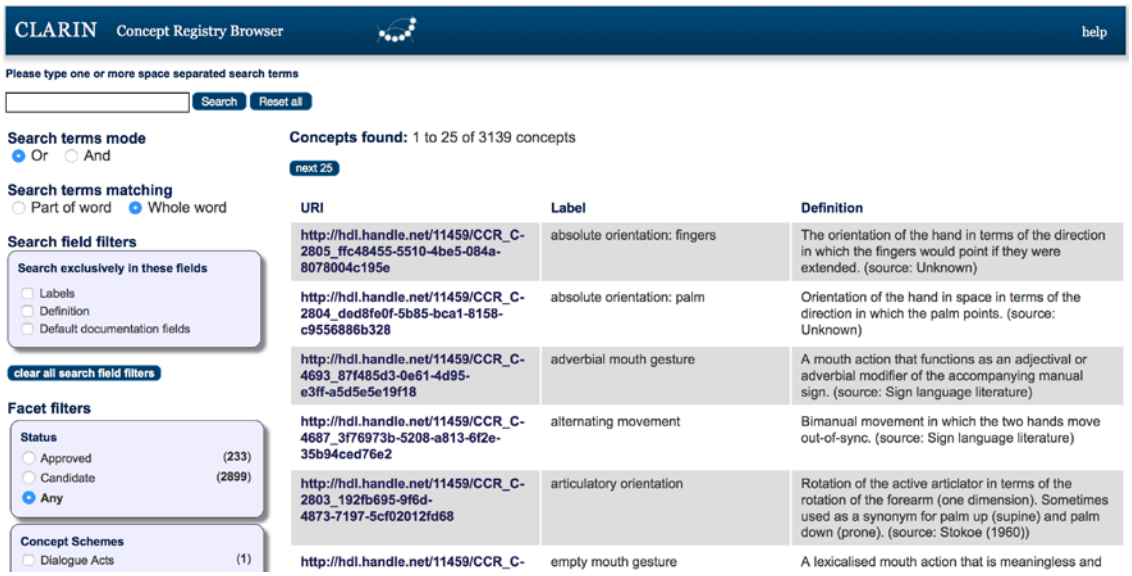

Figure 1. The CCR browser (clarin.eu/conceptregistry)

within CLARIN. With the CCR in place the national CCR content coordinators have teamed up more actively and established procedures around the CCR to fulfil this mission.

To deal with the ISOcat legacy the coordinators are doing a round of clean up with the aim to expire low quality concepts and recommend high quality concepts. Notice that, just like in ISOcat, expired concepts remain accessible, i.e., their semantic descriptions are not lost, but their active usage is discouraged. The main focus is on providing good definitions. A good definition should be "as general as possible, as specific as necessary" and should therefore be:

1. Unique, i.e., not a duplicate of another concept definition in the CCR;
2. Meaningful;
3. Reusable, i.e., refrain from mentioning specific languages, theories, annotation schemes, projects;
4. Concise, i.e., one or two lines should do;
5. Unambiguous.

As far as point 5 is concerned, a concept used in the entry of another concept, should be referred to using its handle. Detailed guidelines are under development by the coordinators and will become generally available in due course. Apart from defining best practice for the coordinator group, such guidelines will benefit users directly, enabling them to issue informed requests to the CCR coordinators (see below).

The changes the coordinators can do to existing concepts are limited, i.e., they should not change the meaning. Only typos, awkward formulations, etc. can be remedied. Otherwise a new concept has to be created, and the original one may be expired.

All the coordinators or their deputies are involved with these changes. In cases where they do not agree a vote might take place and the change will be performed if 70% or more of the coordinators agree. A book keeping of the results of votes is maintained at the CCR section of the CLARIN intranet. The time frame within which the discussions and possibly a vote have to reach a decision is 2 weeks. In the holiday seasons and during the initial startup phase a longer time period can be agreed upon by the coordinators.

Members of the CLARIN community wanting new concepts or changes to existing ones need to contact their national CCR content coordinator. Users from countries with no content coordinator should use the general CCR email address (ccr@clarin.eu) to file their requests. These requests will then be discussed within the national CCR content coordinators forum as described above. Note that in OpenSKOS any changes made to concepts are directly public. Therefore new entries or changes will only be entered after their content has been approved by the content coordinator forum. This procedure will take some time, but should result in better quality concepts, less proliferation and eventually a

higher level of trust of the CCR content than was the case for ISOcat. One can also expect that the need for new concepts will diminish over time due to the CCR covering more and more of the domain.

6 Conclusions and future work

Although CLARIN just started working on the new OpenSKOS-based CLARIN Concept Registry and there is still a lot of ISOcat legacy to deal with, the new registry looks promising. Our feeling is that it will be able to provide a more sustainable and higher quality semantic layer for CMDI. An important lesson from the ISOcat experience is that technology is not always the main problem, although a complicated data model or interface never helps. What we do believe in, is establishing robust yet simple management procedures, as outlined in Section 5. These rely on good teamwork in the national CCR content coordinators forum, together with active involvement of the user community.

Acknowledgements

The authors like to thank the national CCR content coordinators forum for the fruitful discussions on the procedures around the CCR. They also like to thank the Max Planck Institute for Psycholinguistics, CLARIN-NL and the Meertens Institute for their support to realize a smooth transition from ISOcat to the CCR.

Reference

[Broeder et al. 2012] Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. *CMDI: a Component Metadata Infrastructure*. Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop.

[Broeder et al. 2014] Daan Broeder, Ineke Schuurman, and Menzo Windhouwer. 2014. *Experiences with the ISOcat Data Category Registry*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

[Brugman and Lindeman 2012] Hennie Brugman and Mark Lindeman. 2012. *Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service*. Proceedings of the Describing Language Resources with Metadata workshop (LREC 2012), Istanbul, Turkey.

[Durco and Windhouwer 2013] Matej Durco and Menzo Windhouwer. *Semantic Mapping in CLARIN Component Metadata*. 2013. In E. Garoufallou and J. Greenberg (eds.), Metadata and Semantics Research (MTSR 2013), CCIS Vol. 390, Springer.

[ISO 12620 2009] ISO 12620. *Specification of data categories and management of a Data Category Registry for language resources*. 2009. International Organization for Standardization, Geneve.

[Windhouwer 2012] Menzo Windhouwer. *RELcat: a Relation Registry for ISOcat data categories*. 2012. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association (ELRA), Istanbul, Turkey.

[Wright et al. 2014] Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman and Daan Broeder. 2014. *Segueing from a Data Category Registry to a Data Concept Registry*. Proceedings of the 11th international conference on Terminology and Knowledge Engineering (TKE 2014), Berlin, Germany.