# CLARIN Concept Registry: the new semantic registry replacing ISOcat

**First Author**
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

**Second Author**
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

## Abstract

up to 200 words [to be done]
up to 200 words [to be done]
up to 200 words [to be done]
up to 200 words [to be done]

## 1 Introduction

In 2008, the ISO TC37 Data Category Registry (DCR) was created in the form of a database (ISOcat). Its use by various communities, including CLARIN, has grown over the years. However, feedback from (CLARIN) users, coupled with changes in ISO standardization procedures, necessitated a review of the system and operational framework with an aim towards improving usability (cf Broeder et al (2014) and Wright et al (2014)). In the end it was decided that the CLARIN community would no longer make use of ISOcat. Instead, the CLARIN Concept Registry (CCR) has been developed, which is an OpenSKOS registry (cf. Brugman and Lindeman 2012).

Since early 2015, the CLARIN community no longer uses ISOcat as data category registry as the latter was considered to be too complex, asking for more data than necessary for CLARIN purposes. Moreover, it was an open registry, one of the consequences being a proliferation of data.

ISOcat does still exist, and ISO TC37 is planning the next DCR generation. But at the Meertens Institute (Amsterdam), the CLARIN Concept Registry (CCR) has been built, suited to our (CLARIN) needs. The ISOcat entries we need as the CLARIN community are exported into CCR, where they show up in a modified, i.e., simplified, way. Exported are at least all DCs related to CMDI, plus those the national CLARIN groups wanted to be included as they are relevenant for their work. New entries can be added, but in a more controlled way: everybody has read-access, but only the national CCR coordinators can insert new entries, meaning that (proposals for) new entries should be passed on to them. This way we want CCR to become a registry with high-quality content.
The Component Registry now also uses CCR instead of ISOcat.[1]

In this paper we will have a look at CCR in order to see whether the obstacles mentioned in the papers mentioned above ((Broeder et al (2014); Wright et al (2014)), but also those in other ones, like Patejuk and Przepiórkowski (2010) and Zinn at al (2012) have been addressed properly. But first the current content of CCR is described.

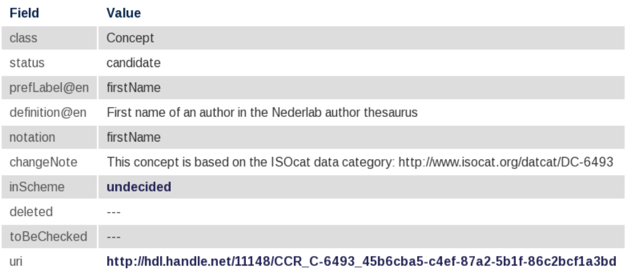
## 2 Content of CCR

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 mentioned to be worthwhile for a specific CLARIN

for review submission

[1] The ISOcat2CCR tool and mapping files that supported this update is available for download and can be used to also update other resources to use CCR handles instead of ISOcat PIDs.

group were inserted in CCR.[2] 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 (https://openskos.meertens.knaw.nl/ccr/browser/index.php, June 2015)). Which entries originate in ISOcat can still be detected, and also what its PID was. Even more, the essential parts of this PID are incorporated in the new URI (see Fig. 1).



| Field | Value |
|---|---|
| class | Concept |
| status | candidate |
| prefLabel@en | firstName |
| definition@en | First name of an author in the Nederlab author thesaurus |
| notation | firstName |
| changeNote | This concept is based on the ISOcat data category: http://www.isocat.org/datcat/DC-6493 |
| inScheme | **undecided** |
| deleted | --- |
| toBeChecked | --- |
| uri | **http://hdl.handle.net/11148/CCR_C-6493_45b6cba5-c4ef-87a2-5b1f-86c2bcf1a3bd** |

Figure 1: Screenshot CCR browser after selecting an URI

Although in the remainder of this paper the improvements in CCR with respect to ISOcat will be highlighted, ISOcat also had some nice features we wanted to maintain in the new registry:

1. public guest access (read-only)
2. edit access for selected group of people
3. editor to facilitate the creation of new entries and to update exiting ones[3]
4. name of the creator of an entry displayed (enabling targeted questions)
5. facility to supersede and deprecate entries

As far as feature 3 above is concerned: although in ISOcat not everybody was allowed access as expert user,[4] far too many people were. In CCR edit access is given to a limited set of people (the national CCR coordinators). As a personal (academic) shibboleth is needed for login, we trust edit access is not offered to other people as is known to have happened in ISOcat.
With respect to feature 5, a 'supersededBy' relation is not yet available in OpenSKOS (and thus not in CCR). It is on the wishlist for OpenSKOS 2.0.

## 3 The main issues concerning ISOcat, and the way these are handled in CCR

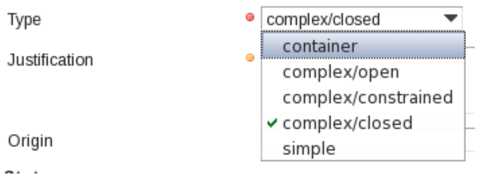
### 3.1 Openness

In ISOcat it was rather easy to become a registred user, and thus getting the possibility to change entries and to insert new ones. In some sense this was too easy, as it led to a number of entries of disputable quality. One of the consequences being that people behind some standards, even ISO ones refused to take the burden of having their standard included.
Within CCR everybody can consult the database, but only a few people, the national CCR-coordinators and their substitutes, can insert new entries (for larger amounts, bulk upload is made possible), i.e., this registry is a closed one. This way the quality of the entries is kept under control, although there is still the need to deal with quality issues in the concepts imported from ISOcat.[5]

[2] In addition, from the Component Registry other (meta)data used in private and public profiles were taken as well.
[3] There are guidelines for doing so. One of the main tasks of the CCR coordinators is to have them observed.
[4] Expert users were allowed to enter new data categories
[5] Cleaning up these entries being one of the first tasks of the CCR coordinators.

### 3.2 Complexity

In ISOcat lots of details not needed by the CLARIN community were asked for, amongst them nasty issues like `type` (see Fig. 2) and `datatype` (see Fig. 3). The first also caused lots of proliferation, the second was in fact choosen at random.

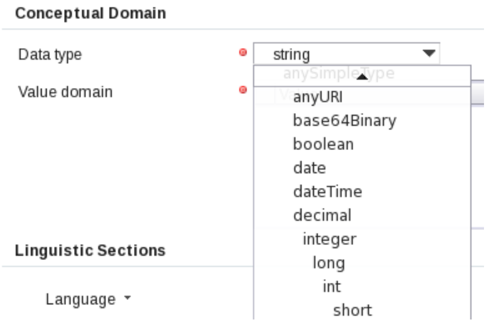


Type ● complex/closed

Justification ●

Origin

Figure 2: Types of DCs



Conceptual Domain

Data type ● string

Value domain ●

Linguistic Sections

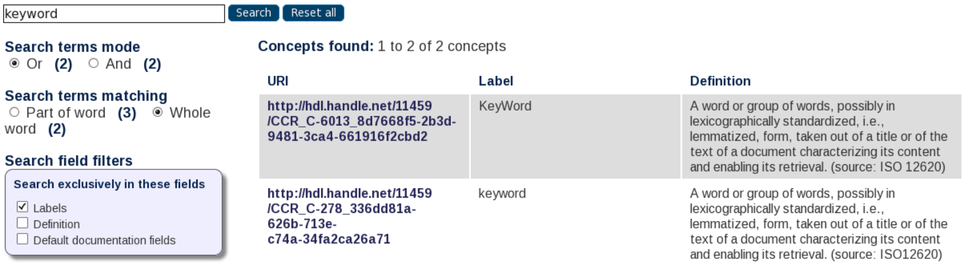
Language

Figure 3: Some 40 more datatypes to go ...

The CCR uses a way less simpler model than ISOcat, i.e., no (DC) types anymore.

### 3.3 Proliferation

The complexity of the ISOcat model sometimes demanded duplicates, e.g., a simple DC /noun/ and a complex DC /noun/ (Fig. 2 above). But there are others reasons as well for duplications, cf. Fig. 4, where in ISOcat different profiles are shown.[6]



keyword [Search] [Reset all]

**Search terms mode**
● Or **(2)** ○ And **(2)**

**Search terms matching**
○ Part of word **(3)** ● Whole word **(2)**

**Search field filters**

**Search exclusively in these fields**
☑ Labels
☐ Definition
☐ Default documentation fields

**Concepts found:** 1 to 2 of 2 concepts

| URI | Label | Definition |
|---|---|---|
| **http://hdl.handle.net/11459/CCR_C-6013_8d7668f5-2b3d-9481-3ca4-661916f2cbd2** | KeyWord | A word or group of words, possibly in lexicographically standardized, i.e., lemmatized, form, taken out of a title or of the text of a document characterizing its content and enabling its retrieval. (source: ISO 12620) |
| **http://hdl.handle.net/11459/CCR_C-278_336dd81a-626b-713e-c74a-34fa2ca26a71** | keyword | A word or group of words, possibly in lexicographically standardized, i.e., lemmatized, form, taken out of a title or of the text of a document characterizing its content and enabling its retrieval. (source: ISO12620) |

Figure 4: Identical definitions, different profiles (ISOcat)/schemes (CCR)

In random order some other causes of (quasi) duplication:

1. public vs private (creator new entry not being aware of existing one)
2. trust/distrust (will the entry remain (semantically) stable?)
3. warnings in the original entry
4. sloppiness (entering an item without looking whether is already does exist)

As far as the first item is concerned, in CCR, all entries are immediately made public.[7] The other issues should be overcome by the closed nature of the new registry, presumed that the coordinators consult each other before accepting new entries.

[6] In the CCR browser shown as different schemes when clicking on the URI. Both entries are to be unified (Currently a cleaning-up operation is done by the coordinators).

[7] This is the standard procedure in OpenSKOS. It implies that the coordinators are to use other ways (like an additional version of CCR with limited read access (both editor and browser)) in the consultation phase. Only when an entry is accepted by them is it transferred to the official CCR. Another option would be to show only accepted entries in the official browser, not the candidate ones.

Note that in CCR when the occasion arises one will also be confronted with two or more definitions for as many entries. For example when linguistic theories/standards disagree.[8] Other duplications may be caused by translation into English: the Dutch concepts /meewerkend voorwerp/ and /belanghebbend voorwerp/ are both translated into /indirect object/ in English.

## 4 Entries in CCR

Few details are asked for in CCR, cf. fig 1, far less than in ISOcat. One of the reasons being, cf section 3.2, that we are now dealing with concepts, rather than with data categories (therefore we can do without the types!). In order to be as reusable as possible, the entries are to represent one (1) concept, rather than a string of concepts. This means, e.g., that complex tags are spit up in its constituent parts which are defined separately. That they can be combined is to be made clear in the manual coming with that tagset, not in the registry itself (solutions also choosen in Patejuk and Przepiórkowski 2010, and recommended in ISOcat, at least for CLARIN.)
Entries showing up in the browser should not be changed in a meaningful way, changing its semantics. Only typos etc can be remedied. Whenever necessary a new enty is to be inserted. In such a case the orinal one may get the status ’expired’.

Definitions should be

1. Unique =¿ no duplicates
2. Meaningful
3. Reusable =¿ refrain from mentioning specific languages, theories, annotation schemes, projects
4. Concise =¿ one or two lines should do
5. Unambiguous

the golden rule being that definitions are

- as general as possible
- as specific as necessary

As far as point 5 is concerned, a concept used in the definition should be liked to its own definition (using its URI). In the future these concepts will be mede clickable.
These guidelines are to be obeyed by the people creating entries. In general, these are not the national CCR coordinators. The task of the latter is just to have the rules obeyed, in cooperation with their co-coordinators, and in the end, to accept (or reject) the entries proposed.

## 5 Conclusions

blablabla

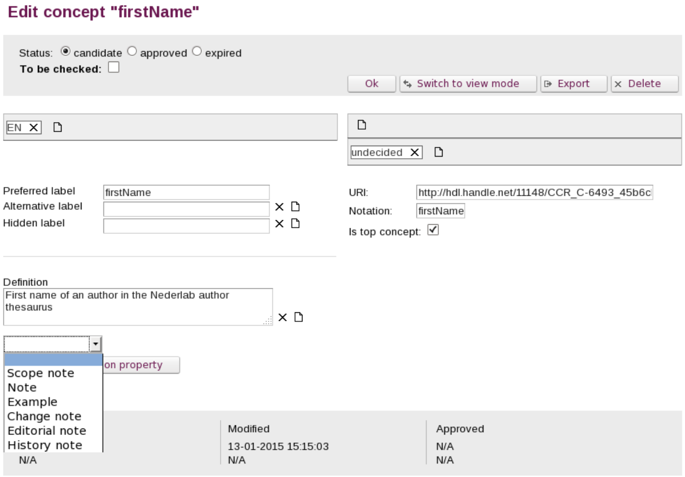
Futere work: relations

## 6 References (still plain text)

Agnieszka Patejuk, Adam Przepiórkowski. ISOcat Definition of the National Corpus of Polish Tagset. In proceedings of the LRT standards workshop (LREC 2010), Malta, May 18, 2010

C. Zinn, C. Hoppermann and T. Trippel. The ISOcat Registry Reloaded. In Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012), Heraklion, Crete, Greece, May 27-31, 2012

[8] In ISO-TimeML, a /state/ is a kind of /event/, whereas in many theories on Tense & Aspect /event/ and /state/ are considered sisters.



Edit concept "firstName"

Status: ● candidate ○ approved ○ expired

To be checked: □

Preferred label: firstName

Alternative label:

Hidden label:

URI: http://hdl.handle.net/11148/CCR_C-6493_45b6c

Notation: firstName

Is top concept: ☑

Definition: First name of an author in the Nederlab author thesaurus

Modified: 13-01-2015 15:15:03, N/A

Approved: N/A, N/A

N/A

Figure 5: Full edit environment

D. Broeder, I. Schuurman, M. Windhouwer. Experiences with the ISOcat Data Category Registry. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association (ELRA), Reykjavik, Iceland, May 28-30, 2014.

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman and Daan Broeder (2014) Segueing from a Data Category Registry to a Data Concept Registry. In Proceedings of the 11th international conference on Terminology and Knowledge Engineering (TKE 2014), Berlin, June 19-20, 2014

Brugman, H. and Lindeman, M. (2012). Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. Istanbul, Describing Language Resources with Metadata workshop