

Content of the Data Category Registry

Data Category Registry (DCR)

- Set of Data Categories (DCs) describing, for example, linguistic annotation schemes, metadata, at several levels/**thematic domains**
 - Morphosyntax
 - Syntax
 - Terminology
 - ...
- **ISOcat** as a whole is a DCR

Thematic Domains (TD) and Groups

- TDs Contain Data Category Selections (DCS)
 - Data Categories within a specific domain (like everything for terminology),

Apart from TDs as reflected by ‘Thematic Views’ in browser pane: ‘groups’

- A subset of DCs, like all DCs for a specific project
- The latter strategy will be followed in CLARIN-NL , making the definitions used in existing corpora etc available first “private/SHARED”, later “public”

Groups

- Group: morphosyntax
 - Entries for:
 - CGN (Corpus of Spoken Dutch)
 - SoNaR (reference Corpus for Dutch)
 - Finnish project XYZ
 - Austrian ...
 - **ALL** concept used are to be included, those needed in the definitions themselves included!!

Groups

Activate DCS and select the "Change scope" icon.

Change the scope of this Data Category Selection

Groups with read and write access to your/their Data Categories in this Data Category Selection

- CLARIN
- GilAndSueEllen
- KSU_PhD_Students
- OTAN
- Part of Speech
- RELISH
- TBX-Basic
- TBX-DCS
- Terminology
- Set scope to *public* (everyone has read access)
- Make all your DCs in this DCS also *public*

Select a group(s) to work on the DCS. Select "Change".

#	Name	Registration status	Check	Type	Owned by
331	abbreviation	1:0 private	candidate	simple	Wright, S
334	acronym	1:0 private	candidate	simple	Wright, S
1230	adjective	1:0 private	private	simple	Francopo
70	administrative			closed	Wright, S
73	admitted to			simple	Wright, S
1232	adverb			simple	Francopo
149	context			open	Wright, S
164	cross referen			open	Wright, S
165	customer s			open	Wright, S
168	definition			open	Wright, S

ISOcat panes

The screenshot shows the ISOcat web interface with several panes highlighted by red boxes and labels:

- Browser pane:** Located on the left side, showing a file tree structure under 'My Workspace' with folders like 'Private', 'Shared', 'Public', and 'Thematic Views'.
- Thematic view:** A sub-section within the browser pane, showing a list of thematic views like 'Metadata', 'Morphosyntax', 'Semantic Content Representation', 'Syntax', 'Language Resource Ontology', 'Lexicography', 'Language Codes', 'Terminology', 'Multilingual Information', 'Lexical Resources', and 'Lexical Semantics'.
- Folder:** A label pointing to the 'Terminology' folder in the browser pane.
- Active DCS pane:** A table displaying a list of terminology entries with columns for ID, Name, Version, Administration status, Registration status, Check, Type, Owned by, and Scope.
- Add:** A button labeled 'Add' located below the Active DCS pane.
- DC specification pane:** A pane showing details for a specific entry (ID 149), including a URL, key, owner, and scope.
- Detach:** A button labeled 'Detach' located to the right of the DC specification pane.
- Basket pane:** A table at the bottom of the interface showing a list of entries in a 'basket', with columns for ID, Name, Version, Administration status, Registration status, Check, Type, Owned by, and Scope.

The main table in the Active DCS pane contains the following data:

#	Name	Version	Administration st	Registration statu	Check	Type	Owned by	Scope
373	normalizedTerm	1.0	accepted	standard	🚩	open	Terminology	public
374	normativeAuthorizati	1.0	accepted	standard	🚩	closed	Terminology	public
112	normativeDocument	0.0.0	private	candidate	🚩	simple	Sue Ellen Wright	public
1487	nounAdjective	0.0.0	private	candidate	🚩	open	Isabelle KRAMER	public
383	nounClass	1.0	accepted	standard	🚩	closed	Terminology	public
1490	nounNoun	0.0.0	private	candidate	🚩	open	Isabelle KRAMER	public
1484	nounPrepositionNour	0.0.0	private	candidate	🚩	closed	Isabelle KRAMER	public

The DC specification pane shows the following details for entry 149:

- Key: 149
- URL: <http://www.isocat.org/datcat/ISO-DC-149>
- complex/open
- Owner: Terminology
- Scope: public

The Basket pane shows the following data:

#	Name	Version	Administration st	Registration statu	Check	Type	Owned by	Scope
347	abbreviatedForm	U.U.U	private	candidate	🚩	simple	Sue Ellen Wright	public
377	admittedTerm	1.0	accepted	standard	🚩	simple	Terminology	public
149	context	1.0	accepted	standard	🚩	open	Terminology	public

CLARIN-NL/VL project level

- Make DCs for all concepts, for example for ‘participle adjective’

Note:

- in running text: */participle adjective/*
 - in identifier: participleAdjective
 - Make use of camelCase !
 - in Data Element Name: participle adjective
-
- For CLARIN-NL (and VL)
 - Make use of Shared CLARIN-NL group !

Language to be used in DC

- identifier, etc
 - Standard in English
 - ALSO when your tagset is defined in another language!

Your tagset “eigennaam” (Dutch)

=> /proper noun/, properNoun, proper noun

But: names in other languages can be added elsewhere.

- Data Element Name is the literal as used in an application/domain

Definition to be used

- The ideal case:
 - There is already a good, standardized definition available which you can reuse
 - Note: this definition should express what is expressed in your application as well!

Een galopperend paard (a galloping horse)

- participle adjective (key 1595) i.e. an adjective
- WW(od,prenom,zonder), i.e. a verb

Definition to be used (2)

- No perfect/good definition available?
 - In case this should/could be fixed
 - We are to negotiate with the owner and try to have it adapted
 - CLARIN-NL/VL: contact the ISOcat coordinator (Ineke)
 - In case of a more fundamental mismatch, come up with a definition yourself
(like verb vs adjective)

New definition

- New definition?

Guidelines:

- English
- One (1) sentence fragment
 - “Element which is used ...”
 - “STTS tag for attributive adjective”
 - “A noun or adjective denoting...”
- Do not mention the concept to be defined
 - zuInclusion \neq inclusion of zu
 - adjective \neq Adjective

[cont.]

New definition (2)

Guidelines (cont.):

- Provide a generic definition whenever possible
 - Especially when a definition was still lacking, not ‘wrong’
- Mention relevant characteristics
- Start with reference to superordinate concept
 - “Pronoun referring to person” for pers.pronoun
- Be as “language neutral” as possible
 - No exotic languages in definition, unless ...

Other aspects of new entries

- Come up with an (mnemonic) identifier
 - Not necessarily unique (unlike key)
- Justify the new entry: why is it important
 - (not: why is existing entry bad !)
- Origin: BNC, Chomsky, ISO, ...
- Provide set of valid values (if applicable) in Conceptual Domain
 - Gender: masculine, feminine, neutral, other

Other aspects of new entries (2)

Optional:

Note section

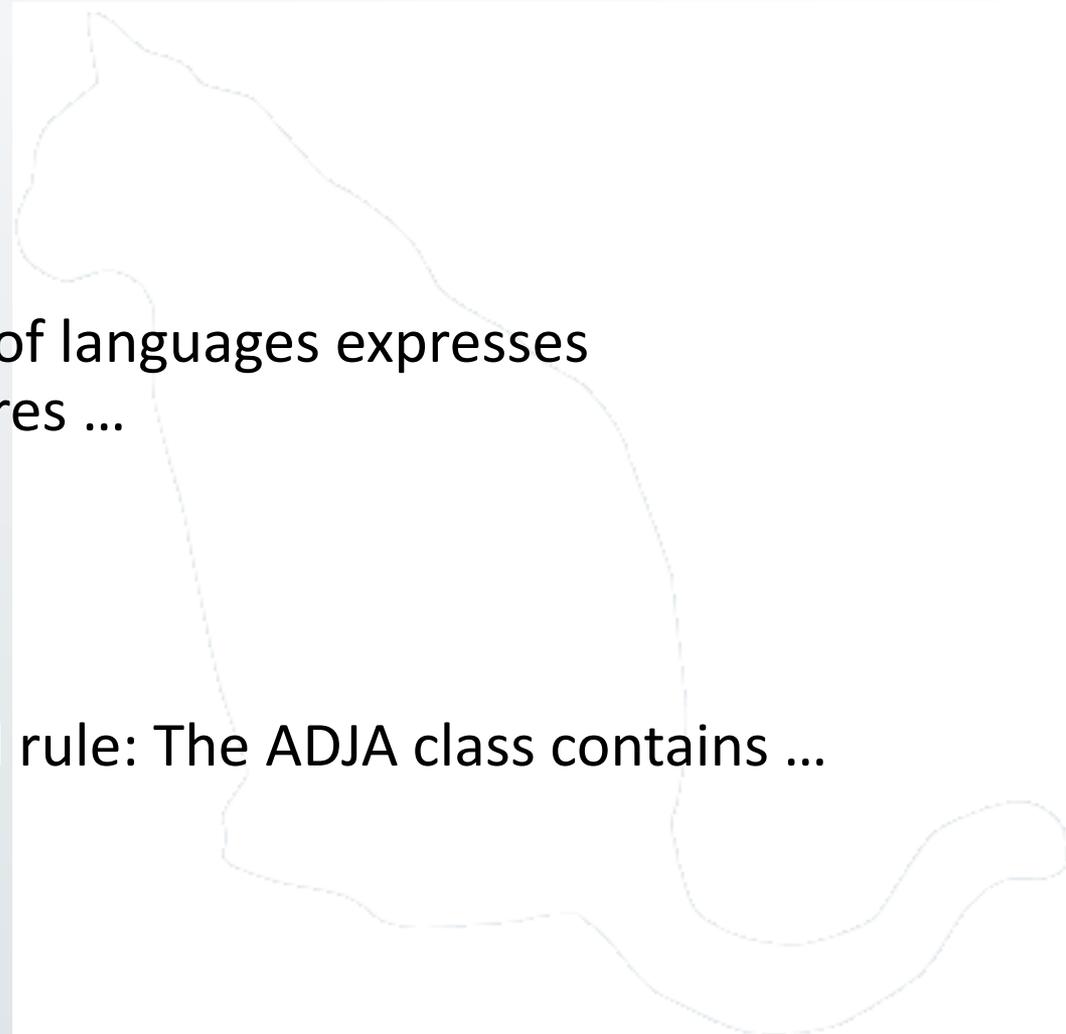
1424 verb

note: A verb in a lot of languages expresses morphological features ...

Explanatory section

2794 ADIA

explanation General rule: The ADJA class contains ...



Profile

- Profile:
 - Indicates which Thematic Domain is involved, this can be more than one:
 - /Part of Speech/ (key 1345):
 - Terminology, morphosyntax
 - In such a case: try to come up with one DC instead of several ones by adding several profiles to one and the same concept
- Failing profile makes it difficult to find a concept
 - /participant age range/

Language/linguistic sections

- Do not confuse these !!
- English/Dutch/Finnish **Language** Section
 - The English section presents the definition (etc.)
 - In the other languages a translation is provided,
Also for the ‘concept’ itself
proper noun => eigennaamThe definition remains in se the same!

Linguistic Section

grammatical gender - 1:0

4. Conceptual Domain

Data Type	string
Profile	Terminology
Value	/feminine/
Value	/masculine/
Value	/neuter/
Value	/otherGender/ (other gender)

5. Linguistic Section

Language English (en)

5.1 Conceptual Domain

Data Type	string
Value	/feminine/
Value	/masculine/
Value	/neuter/

6. Linguistic Section

Language French (fr)

6.1 Example Section

Example	La chien : incorrect
Source	www.atilf.inaif.fr Tifi, 1.GENRE, subst. masc., Gramm

6.2 Conceptual Domain

Data Type	string
Value	/feminine/
Value	/masculine/

7. Linguistic Section

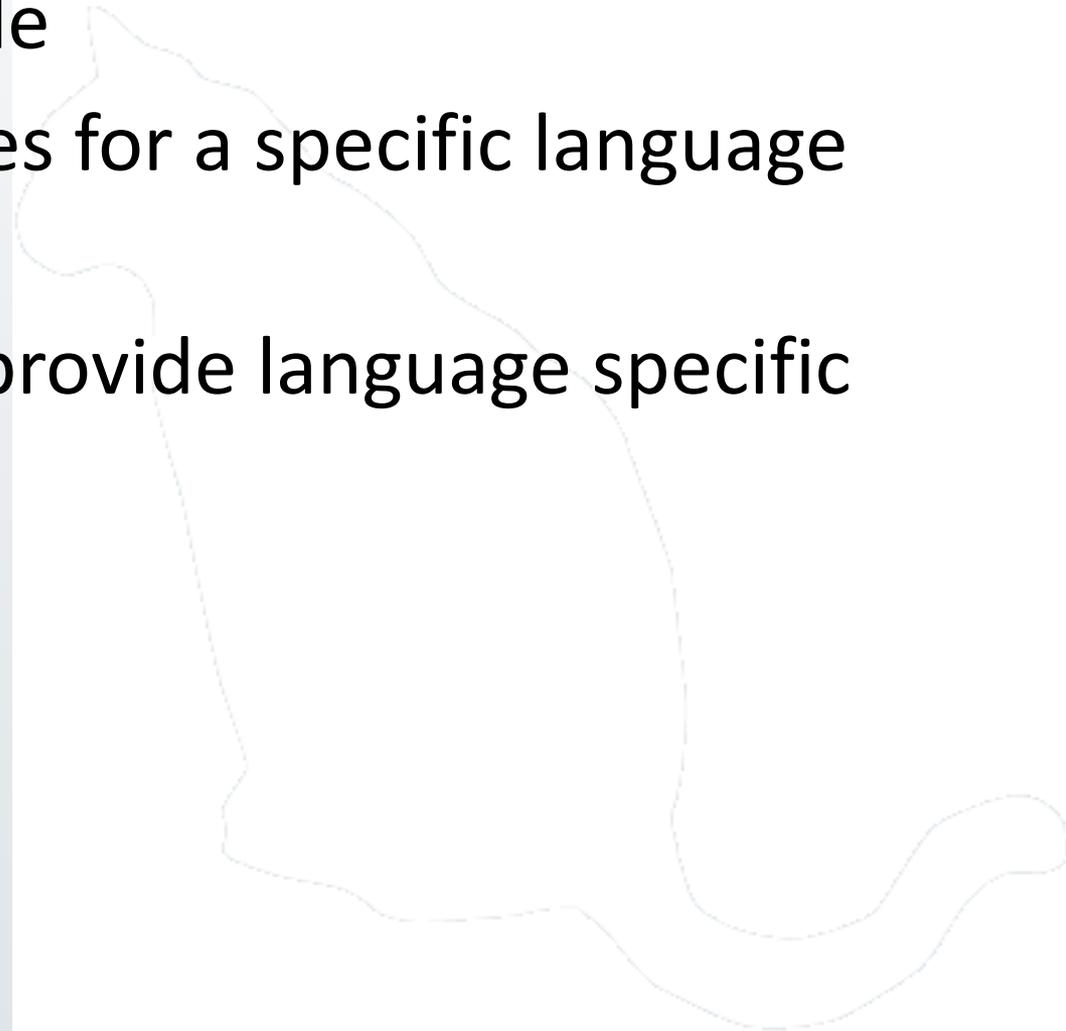
Language German (de)

7.1 Conceptual Domain

Data Type	string
Value	/feminine/
Value	/masculine/
Value	/neuter/

Linguistic Section (2)

- Not always available
- Here relevant values for a specific language can be added
- Also possibility to provide language specific examples



General

- Start in 'Private'
 - CLARIN NL: share your info with the other projects in order to avoid all kinds of problems!
- Public
 - At a later moment the data can be made public.
 - In the end (part of the) DCs may be submitted for standardization (others may be too specific)