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/\textit{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!!
Activate DCS and select the “Change scope” icon.

Select a group(s) to work on the DCS. Select “Change”.
ISOcat panes

Browser pane
Thematic view
Folder
Add
Detach
Active DCS pane
DC specification pane
Basket pane
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 =/= inclusion of zu
  adjective =/= 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 => eigennaam

The definition remains in se the same!
## 4. Conceptual Domain

<table>
<thead>
<tr>
<th>Data Type</th>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile</td>
<td>Terminology</td>
</tr>
<tr>
<td>Value</td>
<td>/feminine/</td>
</tr>
<tr>
<td>Value</td>
<td>/masculine/</td>
</tr>
<tr>
<td>Value</td>
<td>/neuter/</td>
</tr>
<tr>
<td>Value</td>
<td>/otherGender/ (other gender)</td>
</tr>
</tbody>
</table>

## 5. Linguistic Section

<table>
<thead>
<tr>
<th>Language</th>
<th>English (en)</th>
</tr>
</thead>
</table>

### 5.1 Conceptual Domain

<table>
<thead>
<tr>
<th>Data Type</th>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>/feminine/</td>
</tr>
<tr>
<td>Value</td>
<td>/masculine/</td>
</tr>
<tr>
<td>Value</td>
<td>/neuter/</td>
</tr>
</tbody>
</table>

## 6. Linguistic Section

<table>
<thead>
<tr>
<th>Language</th>
<th>French (fr)</th>
</tr>
</thead>
</table>

### 6.1 Example Section

<table>
<thead>
<tr>
<th>Example</th>
<th>La chien : incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td><a href="http://www.atilf.inalf.fr">www.atilf.inalf.fr</a> Tifi, 1.GENRE, subst. masc., Gramm</td>
</tr>
</tbody>
</table>

### 6.2 Conceptual Domain

<table>
<thead>
<tr>
<th>Data Type</th>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>/feminine/</td>
</tr>
<tr>
<td>Value</td>
<td>/masculine/</td>
</tr>
</tbody>
</table>

## 7. Linguistic Section

<table>
<thead>
<tr>
<th>Language</th>
<th>German (de)</th>
</tr>
</thead>
</table>

### 7.1 Conceptual Domain

<table>
<thead>
<tr>
<th>Data Type</th>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>/feminine/</td>
</tr>
<tr>
<td>Value</td>
<td>/masculine/</td>
</tr>
<tr>
<td>Value</td>
<td>/neuter/</td>
</tr>
</tbody>
</table>
Linguistic Section (2)

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