ISOcat
www.isocat.org

Content of the Data Category Registry

ISOcat
www.isocat.org

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

ISOcat
www.isocat.org

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

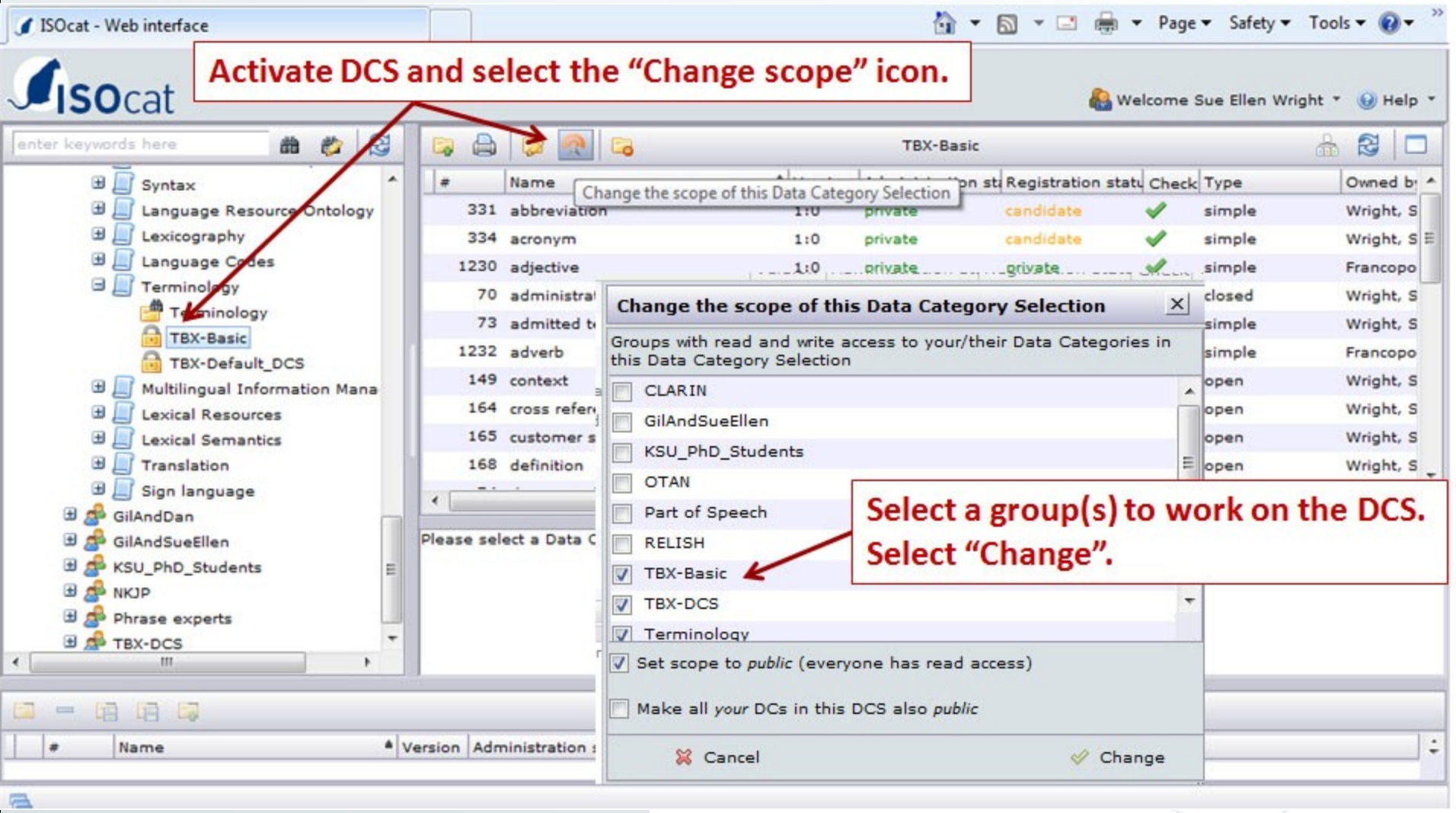
  first "private/SHARED", later "public"

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!!

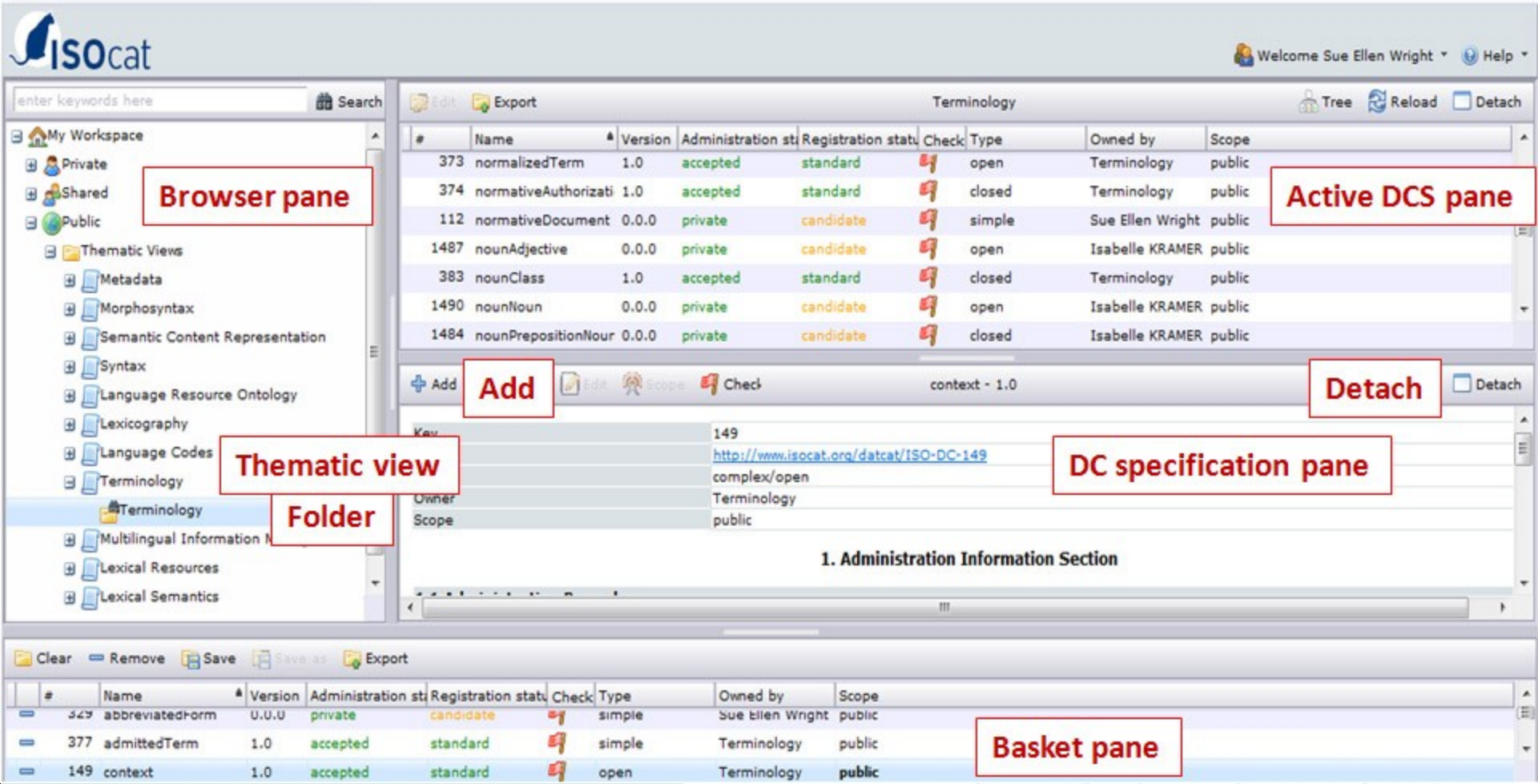
ISOcat
www.isocat.org

Groups



ISOcat
www.isocat.org

ISOcat panes



ISOcat
www.isocat.org

CLARIN-NL/VL project level

- Make DCs for all concepts, for example for ‘particple 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 !

ISOcat
www.isocat.org

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

ISOcat
www.isocat.org

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

ISOcat
www.isocat.org

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)

ISOcat
www.isocat.org

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

ISOcat
www.isocat.org

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/

ISOcat
www.isocat.org

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,

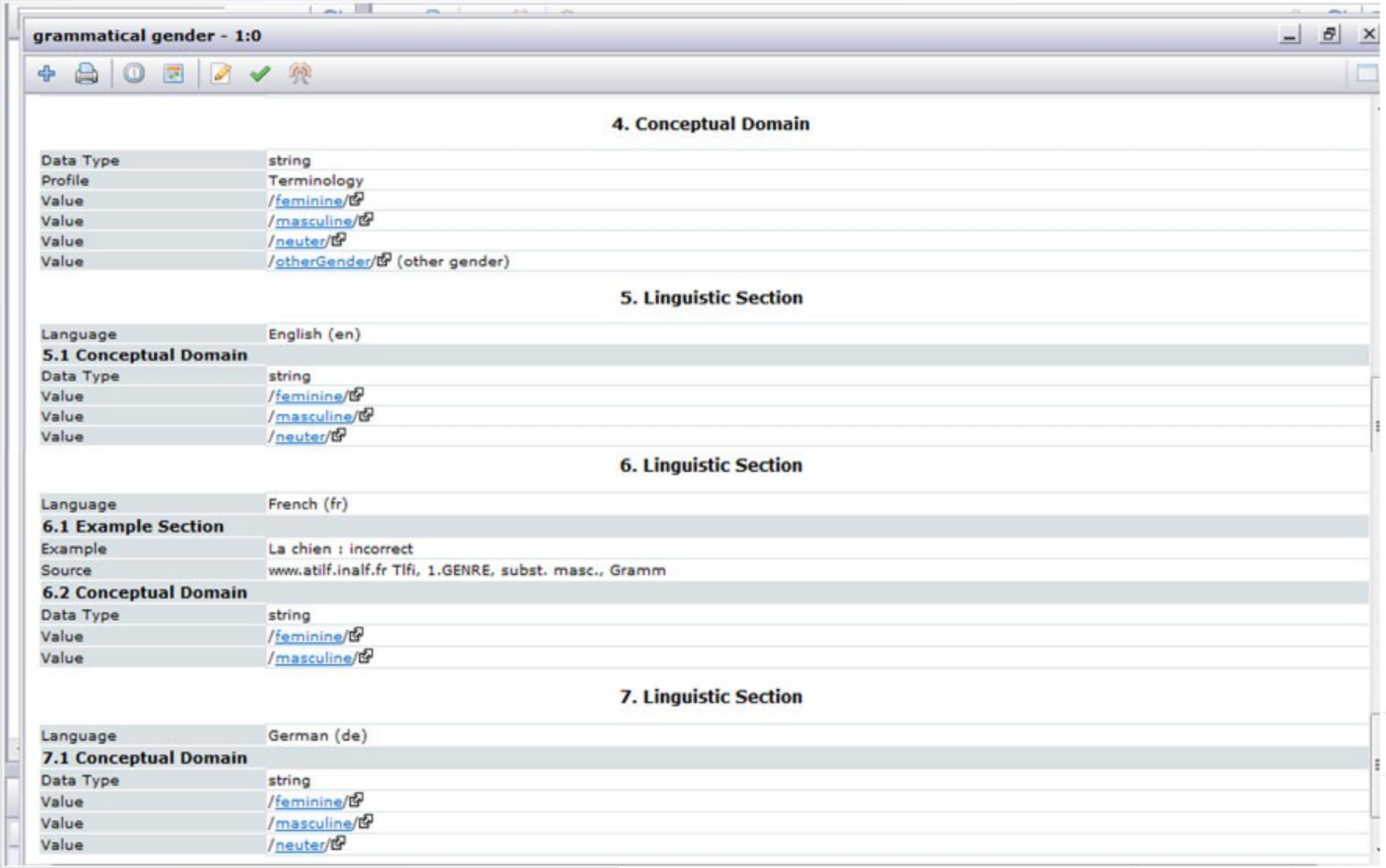
  Also for the 'concept' itself

  proper noun => eigennaam

  The definition remains in se the same!

ISOcat
www.isocat.org

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.inalf.fr Tlfi, 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/

ISOcat
www.isocat.org

Linguistic Section (2)

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

ISOcat
www.isocat.org

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)