ISOcat
www.isocat.org

ISOcat
A short introduction

Marc Kemps-Snijders[a], Sue Ellen Wright[b], Menzo Windhouwer[a]

[a]Max Planck Institute for Psycholinguistics, [b]Kent State University

marc.kemps-snijders@mpi.nl, sellenwright@gmail.com, menzo.windhouwer@mpi.nl

ISOcat
www.isocat.org

ISOcat: a data category registry

- ISO 12620:2009
  - Terminology and other content and language resources — Specification of data categories and management of a Data Category Registry for language resources

Data category

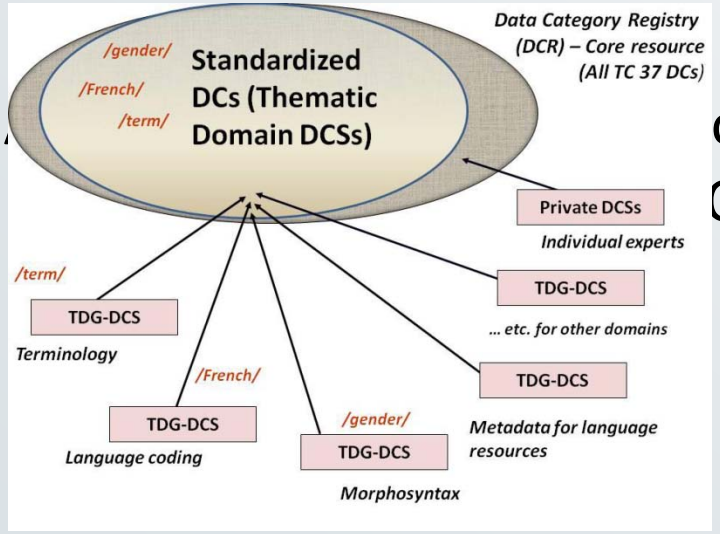
- The result of the specification of a given data field
  - *A data category is an elementary descriptor in a linguistic structure or an annotation scheme.*
- Model consists of 3 main parts:
  - *Administrative part*
    - *Administration and identification*
  - *Descriptive part*
    - *Documentation in various working languages*
  - *Linguistic part*
    - *Conceptual domain(s for various object languages)*

www.isocat.org

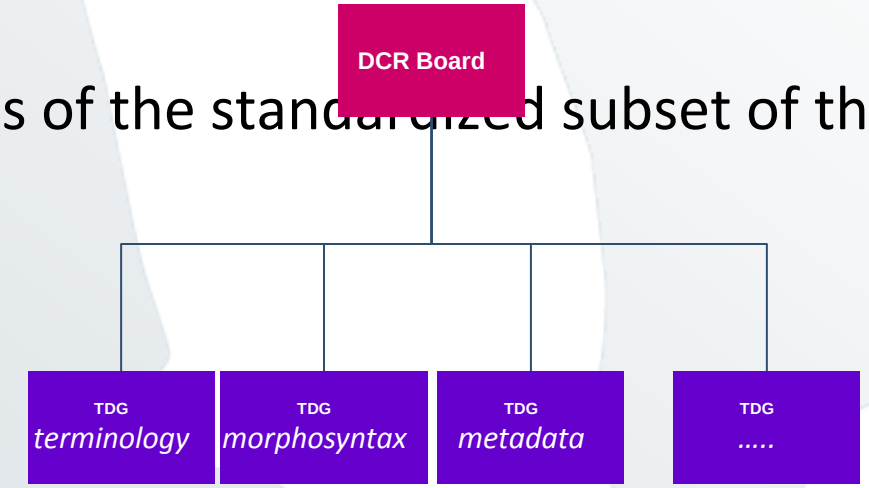
Data Category Registry

- ISOcat is a free service: anyone can access it or register as an expert and create/share his/her own data categories.

- Data categories can be submitted to the standardization process, in which case they are assigned to a Thematic Domain Group which judges it.

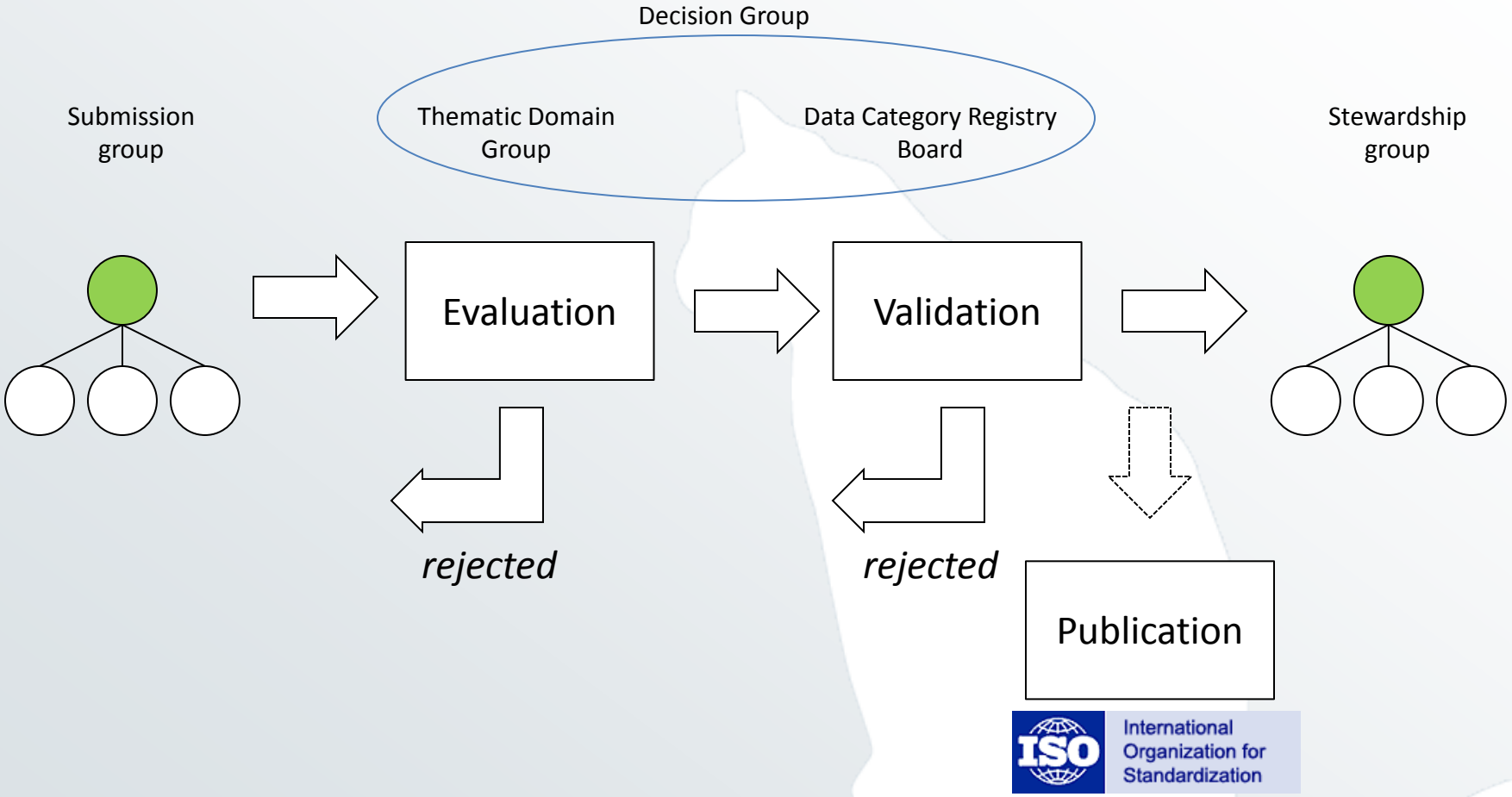


- ots of the standardized subset of the O.



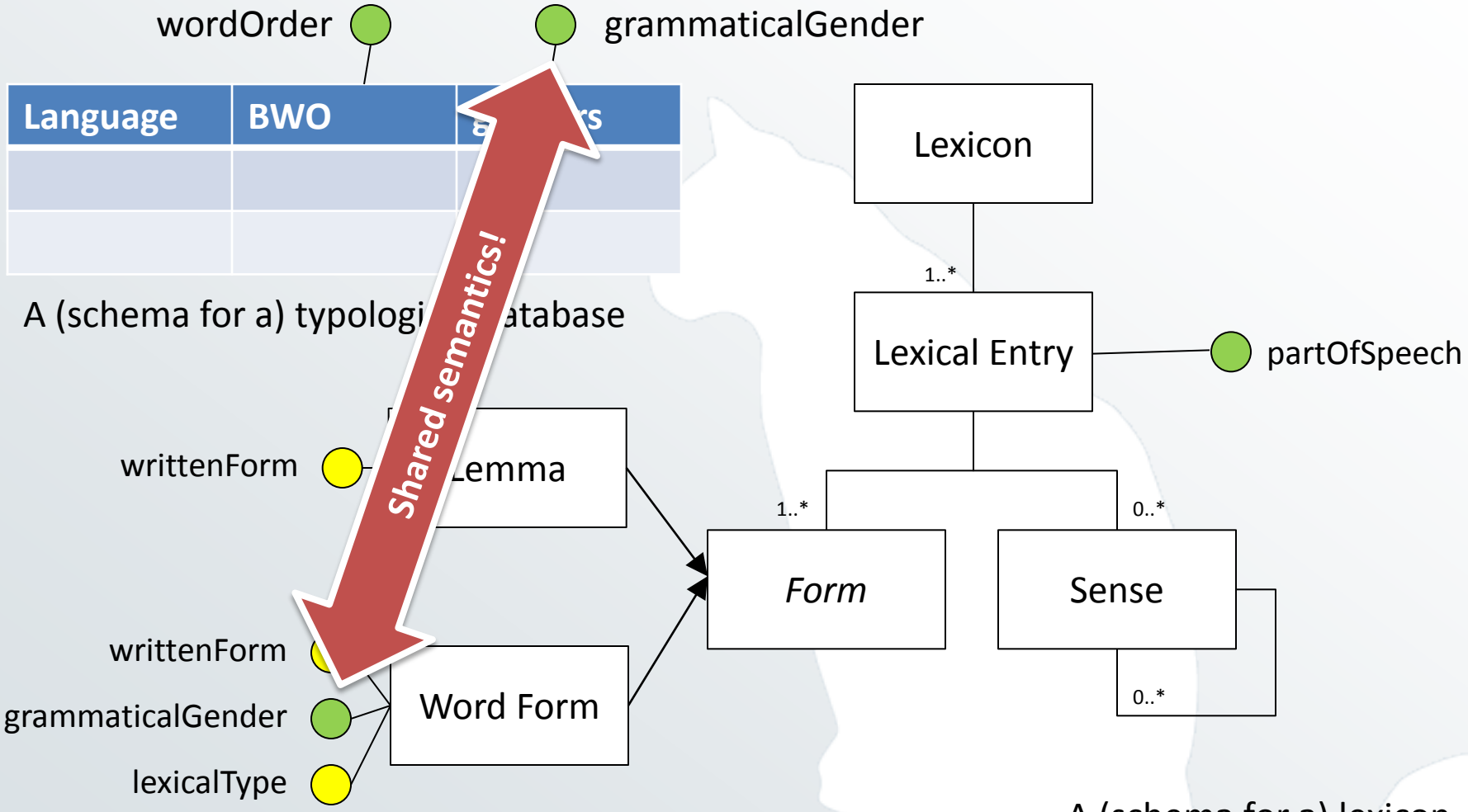
ISOcat
www.isocat.org

Standardization



ISOcat
www.isocat.org

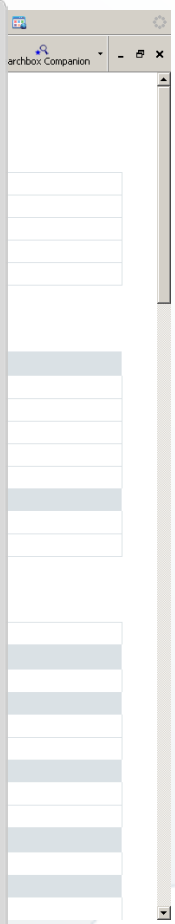
Data categories and linguistic resources



ISOcat
www.isocat.org

Referencing data categories

```
<dcif:dataCategory pid="http://www.isocat.org/datcat/DC-1345" type="complex">
        <dcif:administrationInformation>
                <dcif:administrationRecord>
                        <dcif:identifier>partOfSpeech</dcif:identifier>
                        <dcif:version>0.0.0</dcif:version>
                        <dcif:registrationStatus>candidate</dcif:registrationStatus>
                        <dcif:origin>?</dcif:origin>
                        <dcif:creation>
                                <dcif:creationDate>2004-07-09</dcif:creationDate>
                                <dcif:changeDescription xml:lang="en">
                                        …
                                </dcif:changeDescription>
                        </dcif:creation>
                </dcif:administrationRecord>
        </dcif:administrationInformation>
        <dcif:descriptionSection>
                <dcif:profile>MorphoSyntax</dcif:profile>
                <dcif:languageSection>
                        <dcif:language>en</dcif:language>
                        <dcif:definitionSection>
                                <dcif:definition xml:lang="en">
                                        Term used to describe how a particular word
                                        is used in a sentence.
                        …
                …
        …
</dcif:dataCategory>
```



ISOcat
www.isocat.org

Annotating linguistic resources

- Schema language support for equivalence*:*
  - for example ODD from TEI

```
<elementSpec ident="pos">
        <equiv name="partOfSpeech" uri="http://isocat.org/datcat/ISO-DC-369"/>
       …
</elementSpec>
```

- Annotation using dcr:datcat attribute:
  - for schemas or instances
  - for example RelaxNG schema

```
<rng:element name="partOfSpeech" dcr:datcat="http://isocat.org/datcat/ISO-DC-369" >
        <rng:choice>
                <rng:value dcr:datcat="http://isocat.org/datcat/ISO-DC-370">
                        verb
                </rng:value>
                <rng:value dcr:datcat="http://isocat.org/datcat/ISO-DC-371">
                        noun
                </rng:value>
        </rng:choice>
</rng:element>
```

- XML oriented, is more needed?

www.isocat.org

Data categories as RDF resources

```
:headword
        dcr:datcat <http://isocat.org/datcat/DC-258> ;
        rdfs:label "head word"@en ;
        rdfs:comment "A lemma heading a dictionary entry."@en ;
        rdfs:label "lemma"@nl ;
        rdfs:comment "Het eerste woord van een artikel in een
                woordenboek."@nl .


:partOfSpeech
        dcr:datcat <http://isocat.org/datcat/DC-396> ;
        rdfs:label "part of speech"@en ;
        rdfs:comment "A category assigned to a word based on its grammatical and
                semantic properties."@en .
```

A domain modeling approach:

```
:headword a rdfs:Class .

:partOfSpeech a rdf:Property ;
        rdfs:domain :headword .
```

Alternative approach:

```
:headword a rdfs:Class .

:partOfSpeech a rdf:Class.

:hasPartOfSpeech a rdf:Property ;
        rdfs:domain :headword
        rdfs:range:partOfSpeech.

:noun a partOfSpeech.
```

ISOcat
www.isocat.org

ISOcat status

- ISOcat is under active development:
  - Now:
    - You can access public data categories and selections
    - You can create your own data categories and selections
    - You can share your data categories and selections with others (everyone, or a specified group)
  - In progress:
    - Cleanup of profiles by TDGs
    - Standardization workflow
    - Some social features (forum to discuss specific data categories)
    - Import external 'data category' sets, such as:
      - parts of the ISO Concept Database
      - Dublin Core
      - TEI
  - Future:
    - High availability (mirrors)
    - Relation registry

ISOcat
www.isocat.org

ISOcat workshop

- Utrecht, Thursday March 25, 2010
- Espec ojects
- Signu
- Progr
  - A d
  - A tu
  - Hov es?

Invitation

Send examples of the types of linguistic resources your project wants to annotate with data category references to

isocat@mpi.nl

and we will discuss them at the workshop!

Thank you for your attention!

Visit

www.isocat.org

Questions?

isocat@mpi.nl