ISO 12620 Data Category Registry
An 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

CLARIN-NL ISOcat workshop
Utrecht
2010-03-25

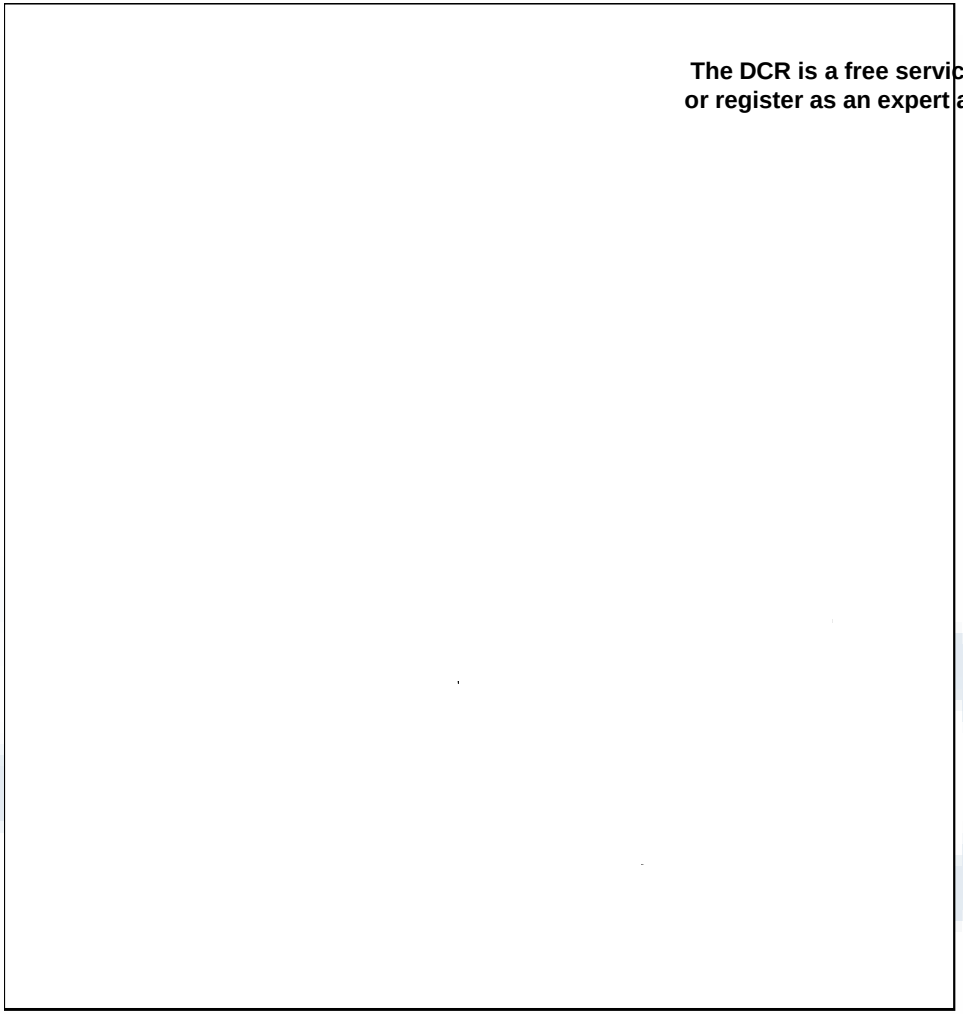
Data category

- 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 and working language*
  - *Linguistic part*
    - *Conceptual domain of object language*

Data category

Administrative part



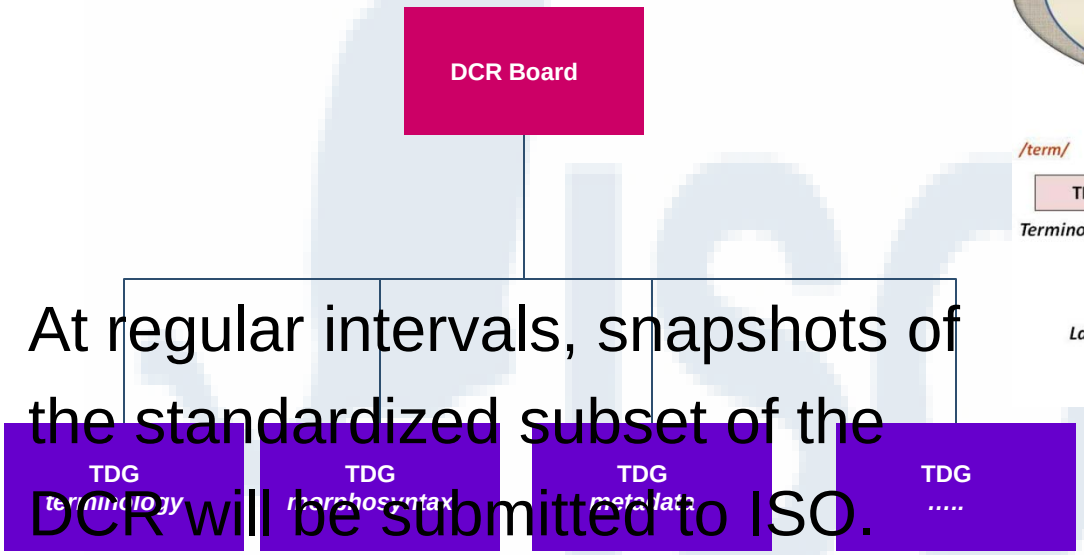
The DCR is a free service: anyone can access it or register as an expert and create/share his/her own data categories.

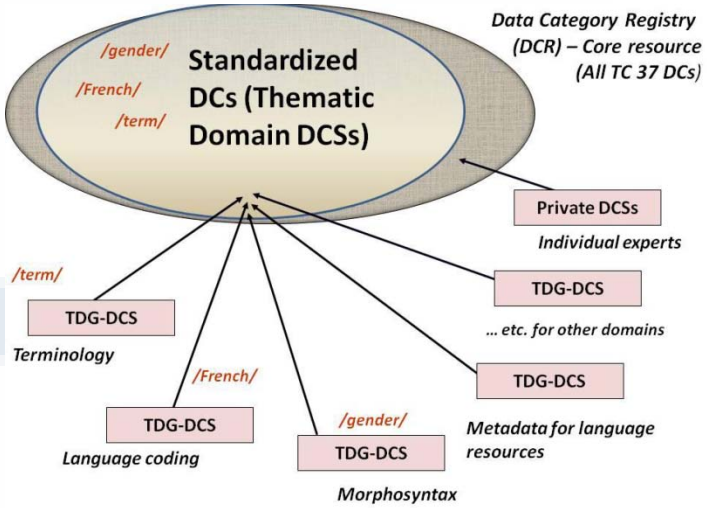
Data categories can be submitted to the standardization process, in which case they are assigned to a Thematic Domain Group which judges it.

Data category

A short excursion into standardization

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

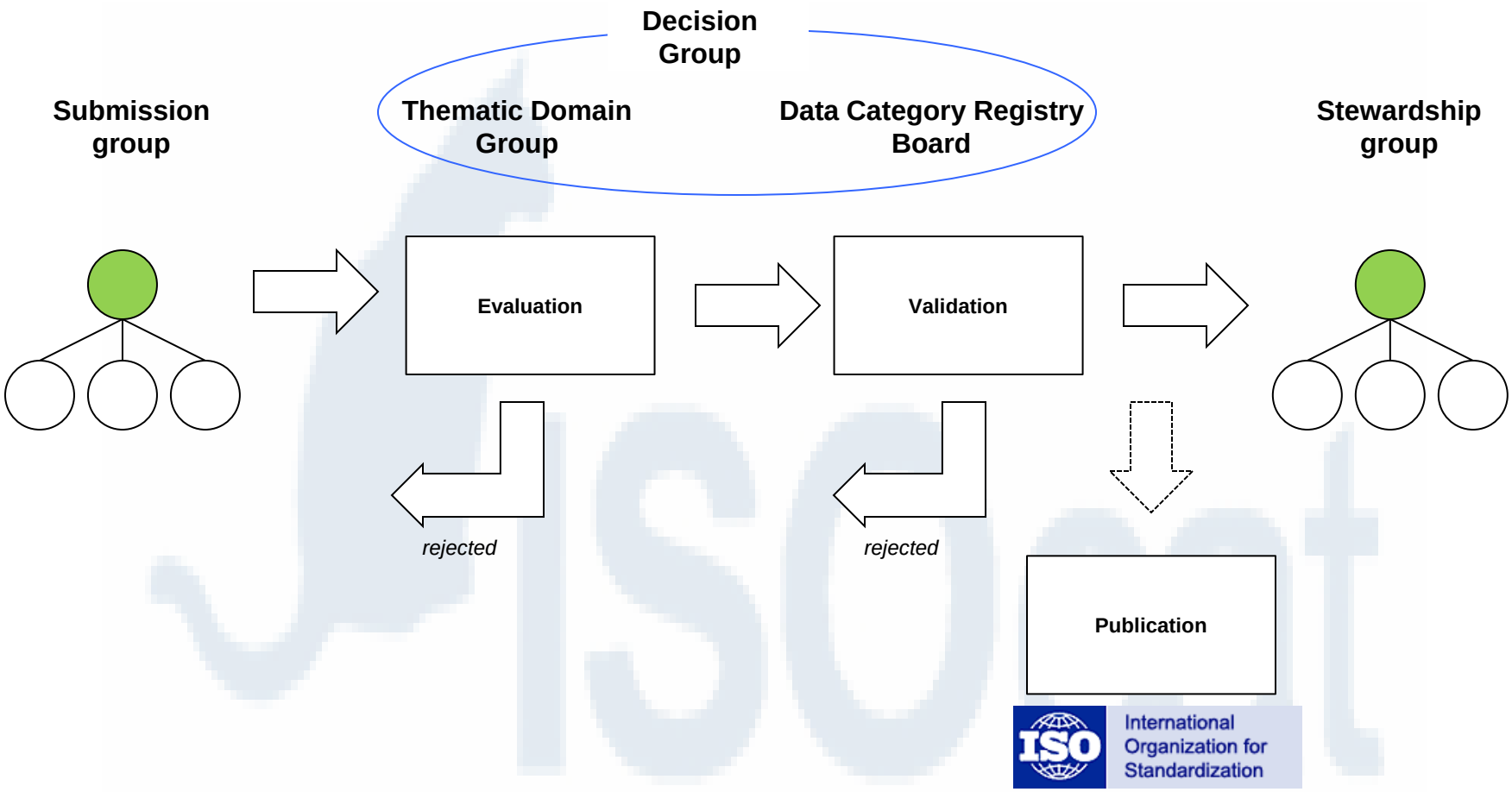




- At regular intervals, snapshots of the standardized subset of the DCR will be submitted to ISO.

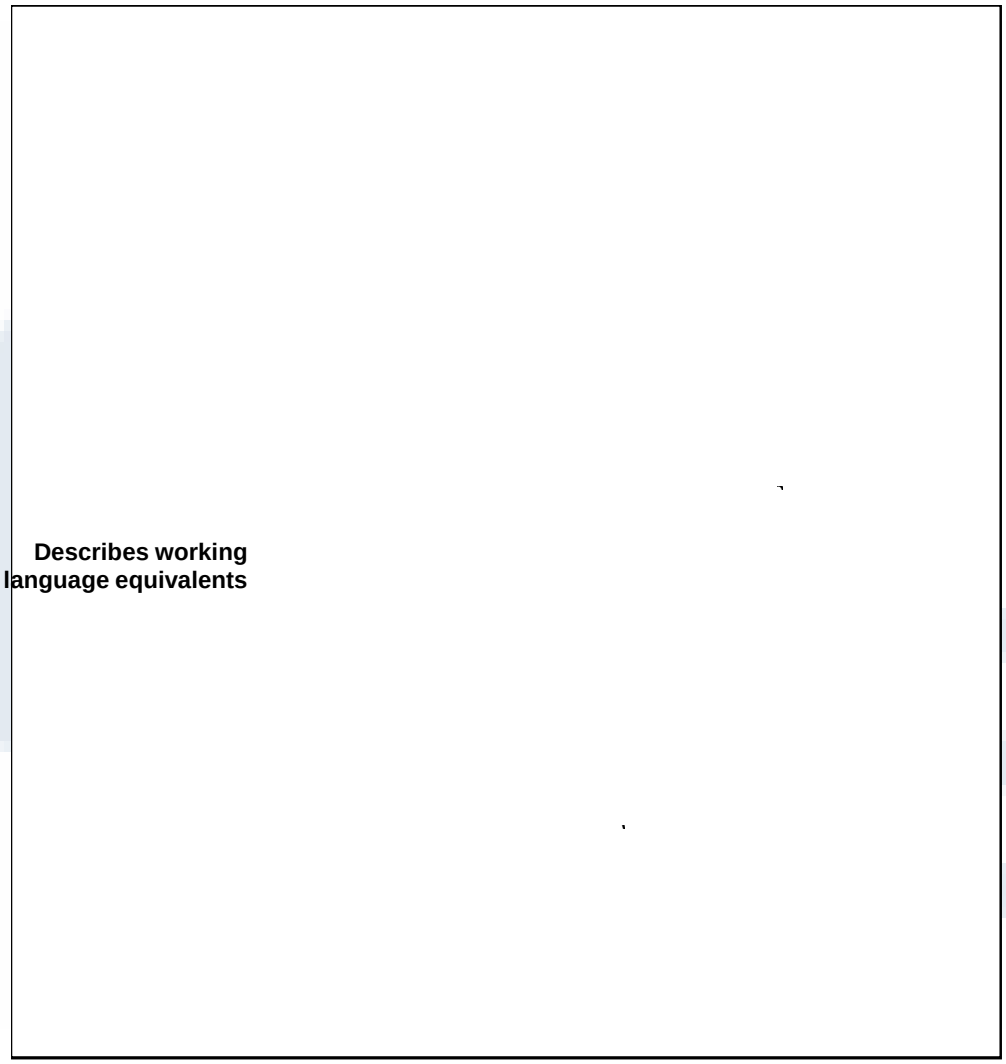
Data category

A short excursion into standardization



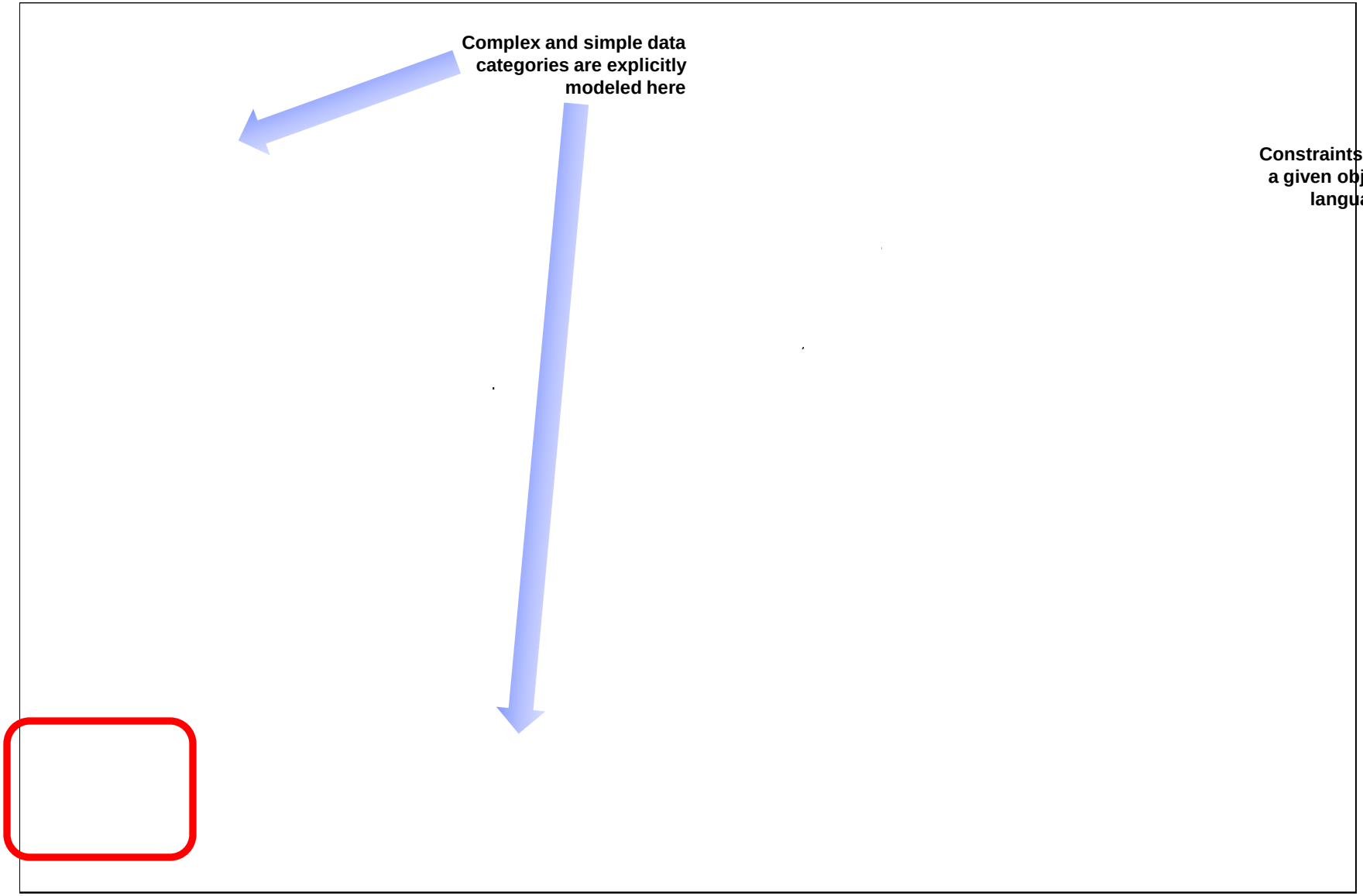
Data category

Descriptive part



Data category

Linguistic part



Data category

Linguistic part (example)

- Data category: */Grammatical gender/*
  - Conceptual domain: */male/*, */feminine/*, */neuter/*
    - *Lists all admissible values for all languages*
  - Linguistic Section
    - Language: fr
    - Value Domain: */male/, /feminine/*
    - *Lists all admissible value for French*

Referencing data categories

DCIF Examp

<dcif:dataCategory pid=

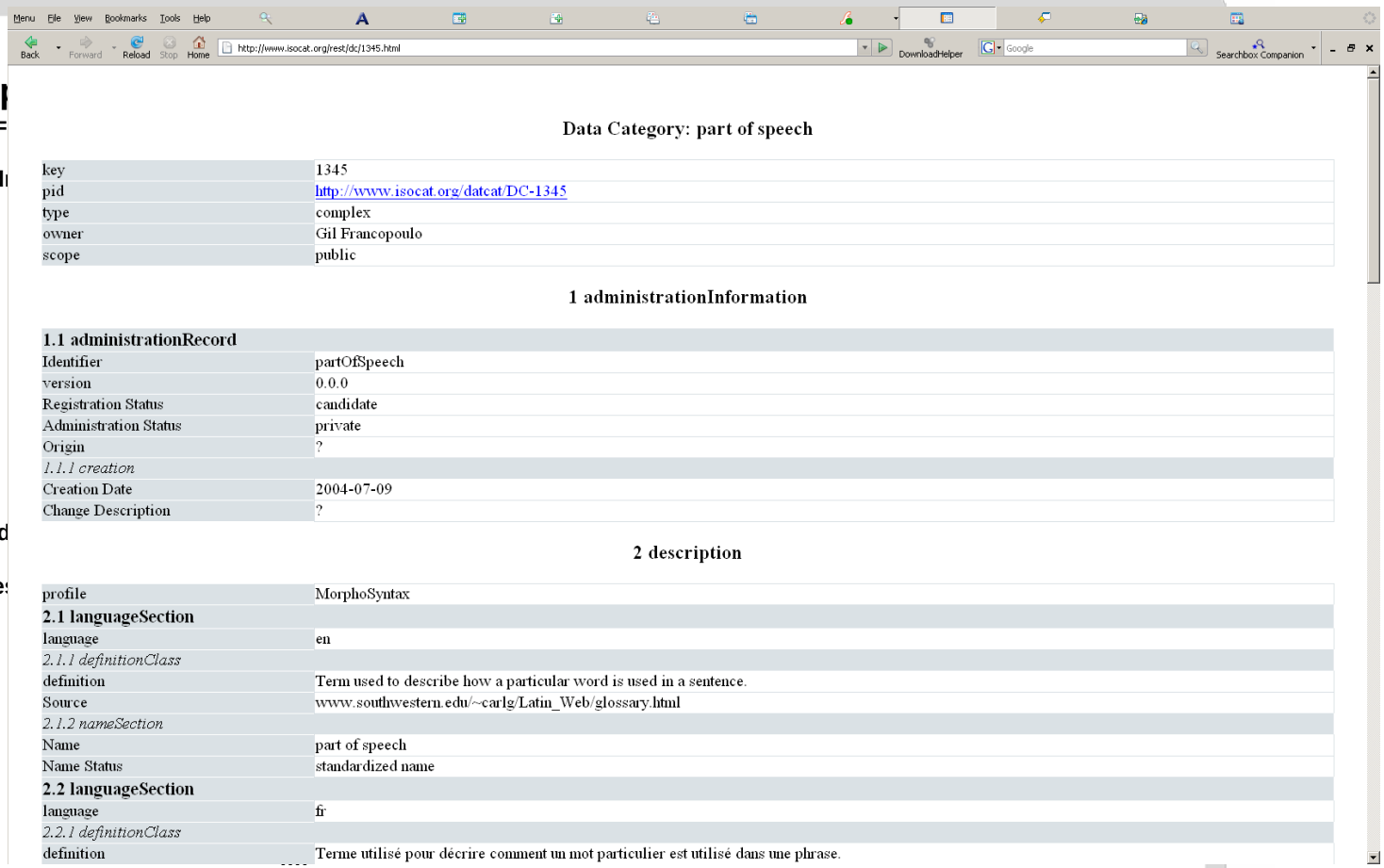
<dcif:ad

</dcif:ad

<dcif:de

...

</</dcif:dataCategory>



Data Category: part of speech

key	1345
pid	http://www.isocat.org/datcat/DC-1345
type	complex
owner	Gil Francopoulo
scope	public

1 administrationInformation

1.1 administrationRecord	
Identifier	partOfSpeech
version	0.0.0
Registration Status	candidate
Administration Status	private
Origin	?
1.1.1 creation	
Creation Date	2004-07-09
Change Description	?

2 description

profile	MorphoSyntax
2.1 languageSection	
language	en
2.1.1 definitionClass	
definition	Term used to describe how a particular word is used in a sentence.
Source	www.southwestern.edu/~carlg/Latin_Web/glossary.html
2.1.2 nameSection	
Name	part of speech
Name Status	standardized name
2.2 languageSection	
language	fr
2.2.1 definitionClass	
definition	Terme utilisé pour décrire comment un mot particulier est utilisé dans une phrase.

Annotating linguistic resources

- Schema language support for equivalence*:*
  - E.g. ODD from TEI

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

- Annotation using dcr:datcat attribute:
  - E.g. RNG 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>
```

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 .
```

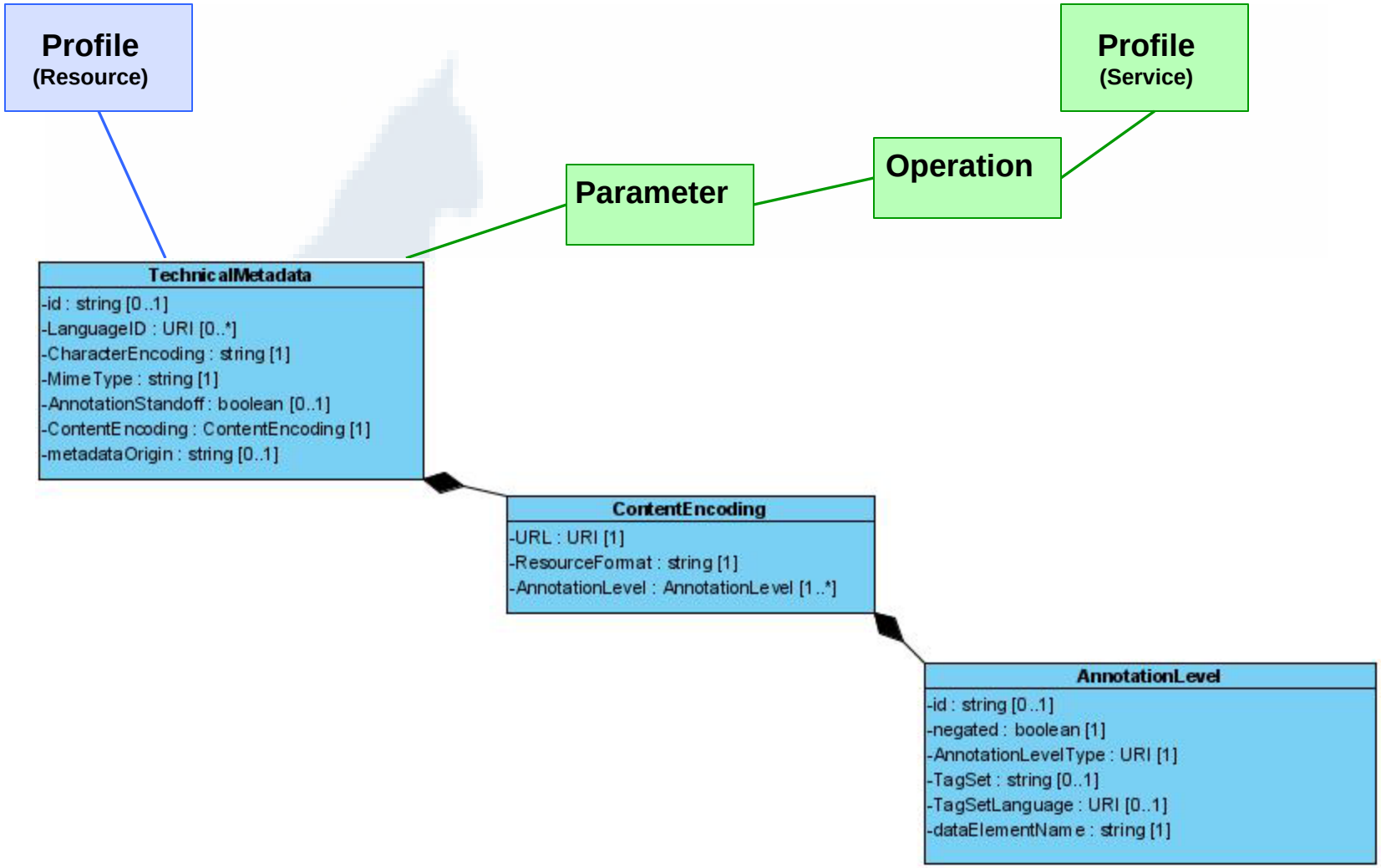
A domain modeling approach:

```
:headword a rdfs:Class .

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

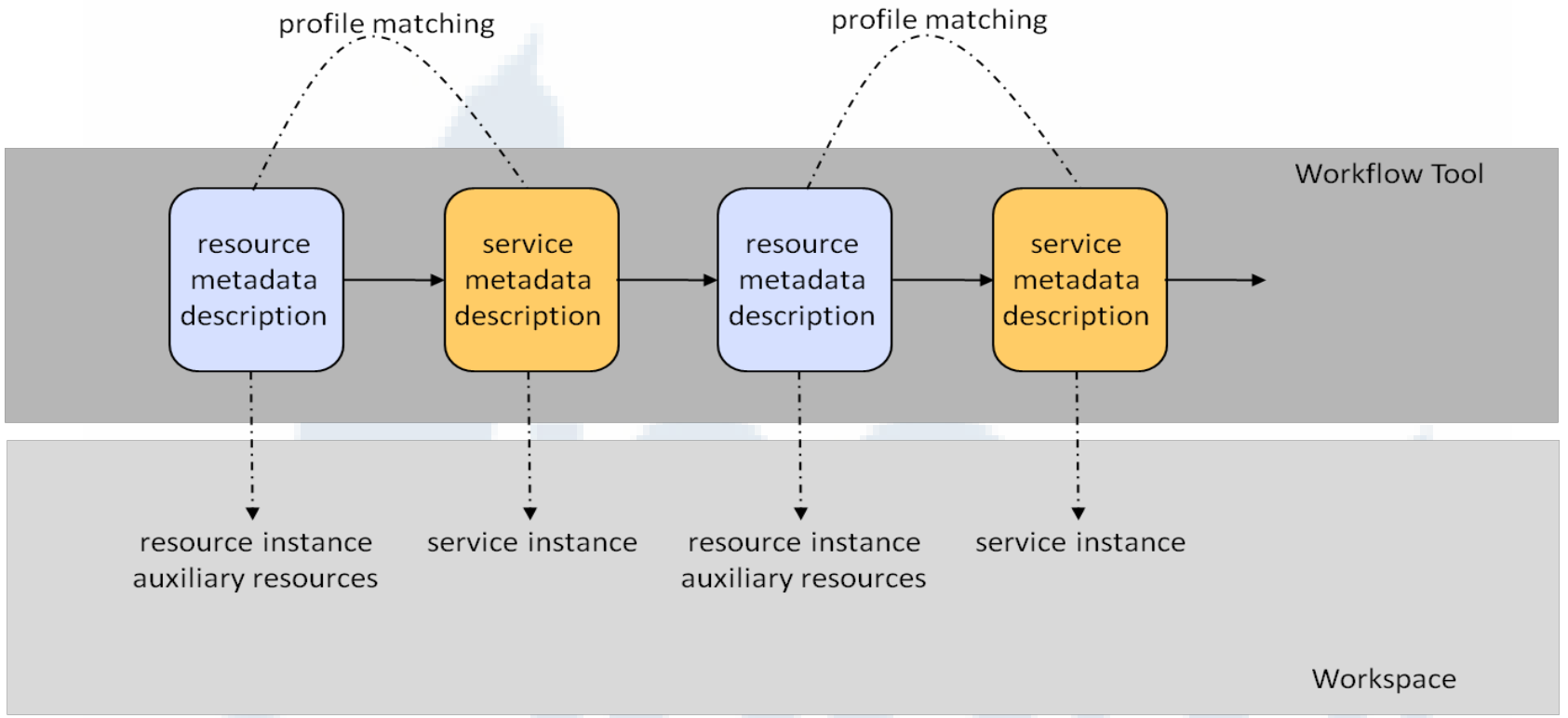
Alternative approach:

```
:headword a rdfs:Class .

:partOfSpeech a rdf:Class.

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

:noun a partOfSpeech.
```

Metadata profiles

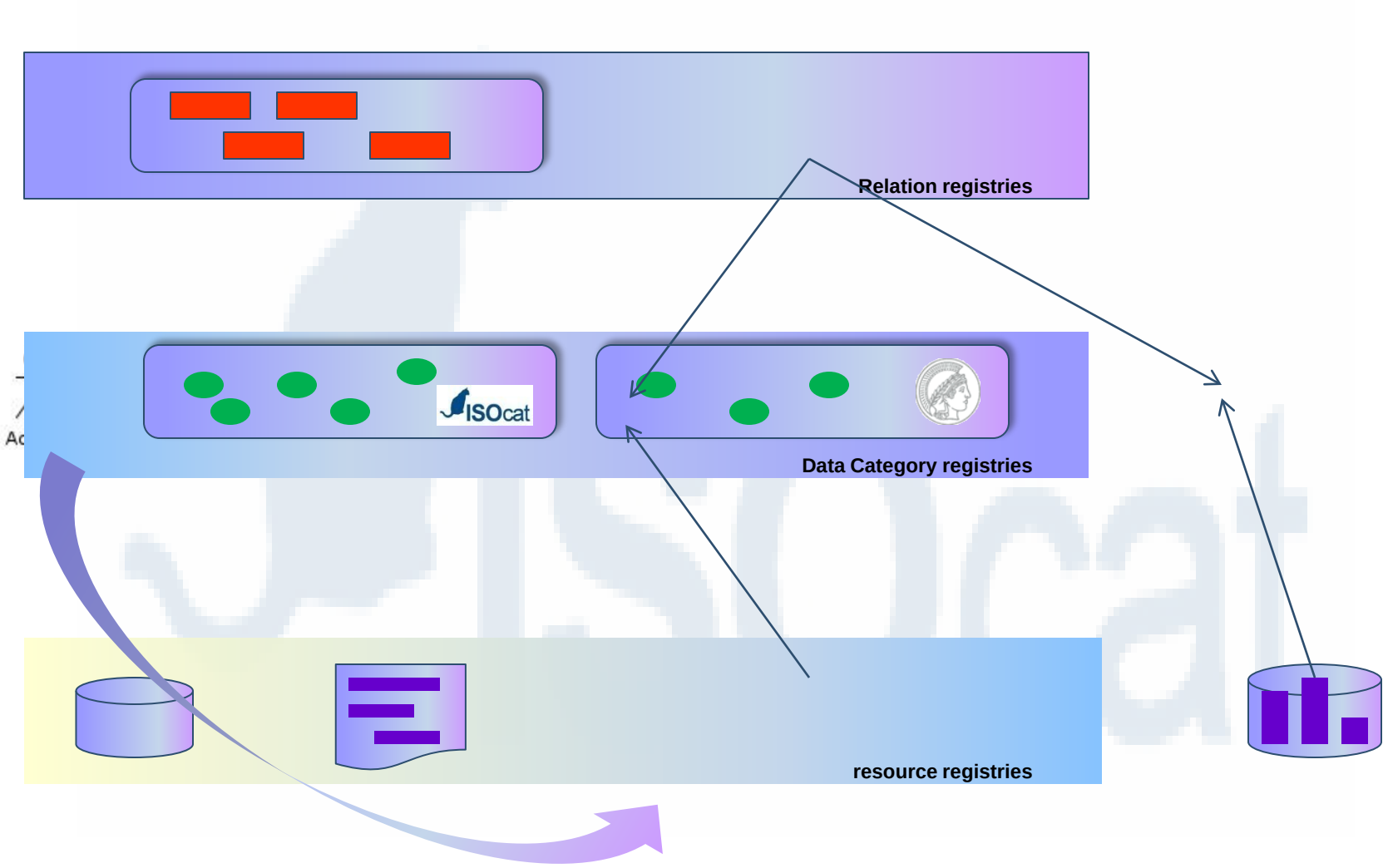


Profile matching



- **metadata of resources and services are used to identify compliance**
- **if not they should suggest converters, etc.**
- **metadata specifications for web services are used for profile matching**

A network of registries



Projects using ISOcat

- ISOcat is at the center of CLARIN-EU
  - Metadata TDG is currently very active here
- In the first project round for CLARIN-NL resources will be annotated using data categories from ISOcat
- D-Spin (CLARIN Germany) project is entering STTS tagset
- GOLD community is involved to populate ISOcat with data categories derived from GOLD

Thank you for your attention