

ISOcat and CMDI

Menzo Windhouwer
MPI for Psycholinguistics
menzo.windhouwer@mpi.nl



Outline

- ISOcat: a Data Category Registry
- The role of data categories in CMDI
- A glimpse of ISOcat
- Status of the metadata profile

ISOcat: a Data Category Registry

- The reference implementation of ISO 12620:2009
 - Terminology and other content and language resources — Specification of data categories and management of a Data Category Registry for language resources
- A data category
 - is the result of the specification of a given data field
 - an elementary descriptor in a linguistic structure or an annotation scheme

Data category specification

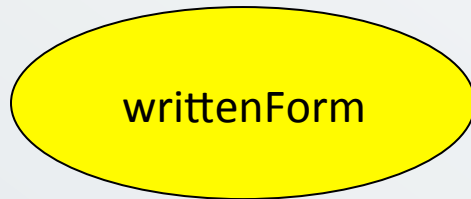
- Administrative part
 - Identifier
 - Version
 - Origin
 - Justification
 - Status
- Descriptive part
 - Names, definitions, examples and explanations in various languages (English is mandatory)
 - Application (domain) specific names
- Conceptual domain
 - Possible values (per profile)
- Linguistic part
 - Examples and explanations for various languages
 - Possible values for various languages

Data Category example

- Data category: */Grammatical gender/*
 - Administrative part:
 - Identifier: grammaticalGender
 - PID: <http://www.isocat.org/datcat/DC-1297>
 - Descriptive part:
 - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
 - French definition: Catégorie fondée (selon la langue) sur la distinction naturelle entre les sexes ou d'autres critères formels.
 - Conceptual domain:
 - Morposyntax conceptual domain: */male/, /feminine/, /neuter/*
 - Linguistic part:
 - French conceptual domain: */male/, /feminine/*

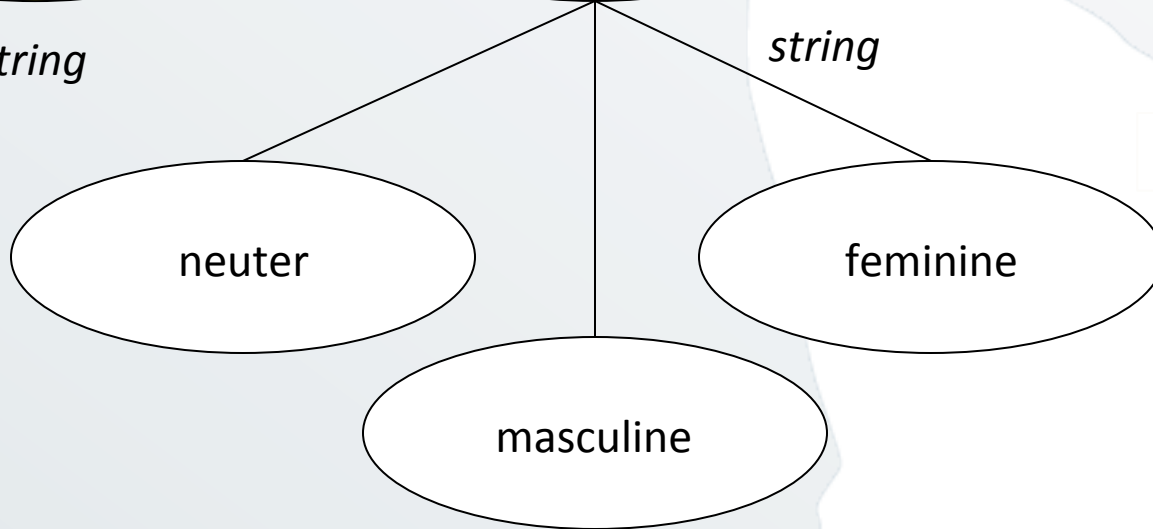
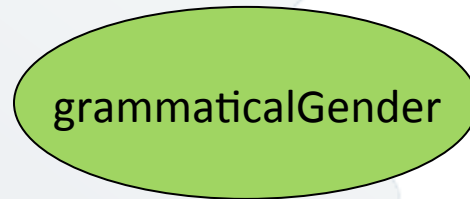
Data Category types

complex: open



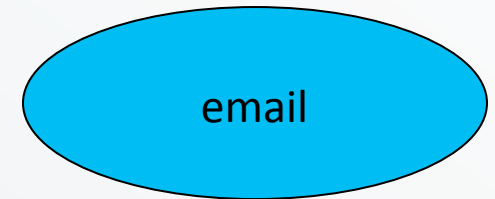
string

closed



string

constrained



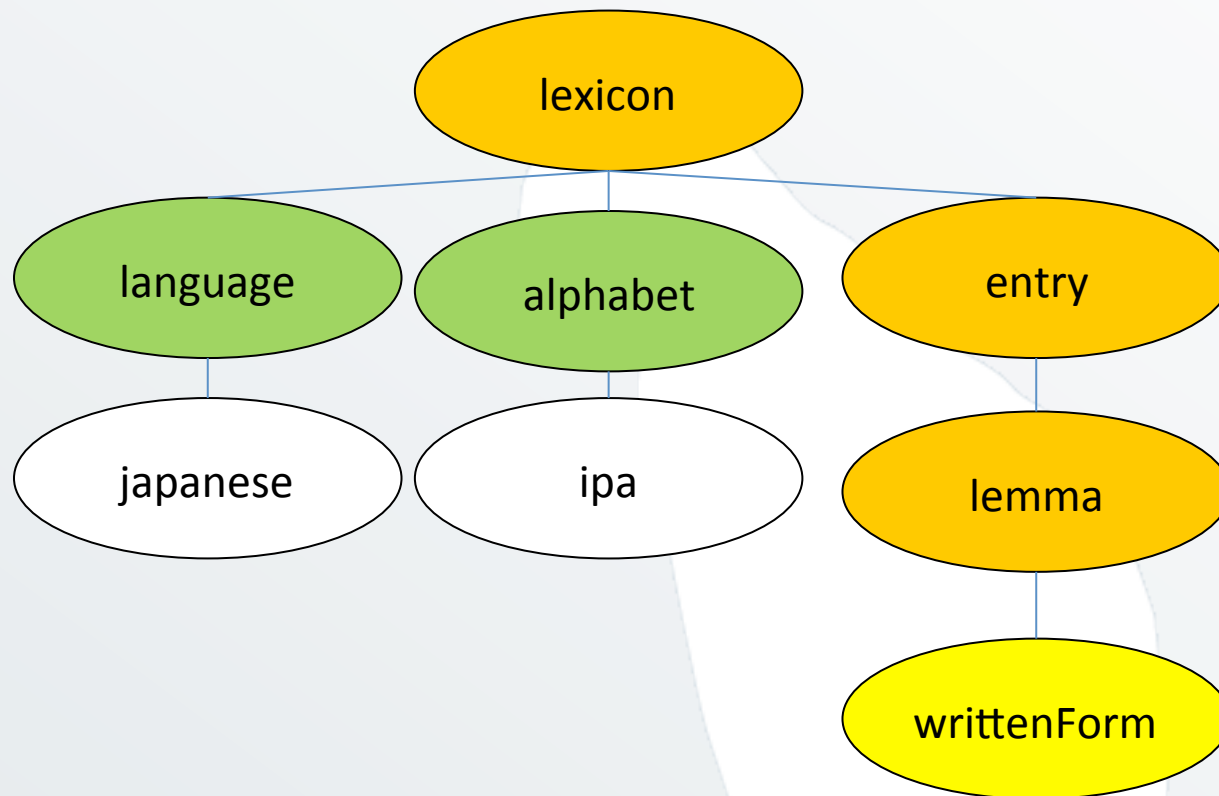
string

Constraint: .+@.+

simple:

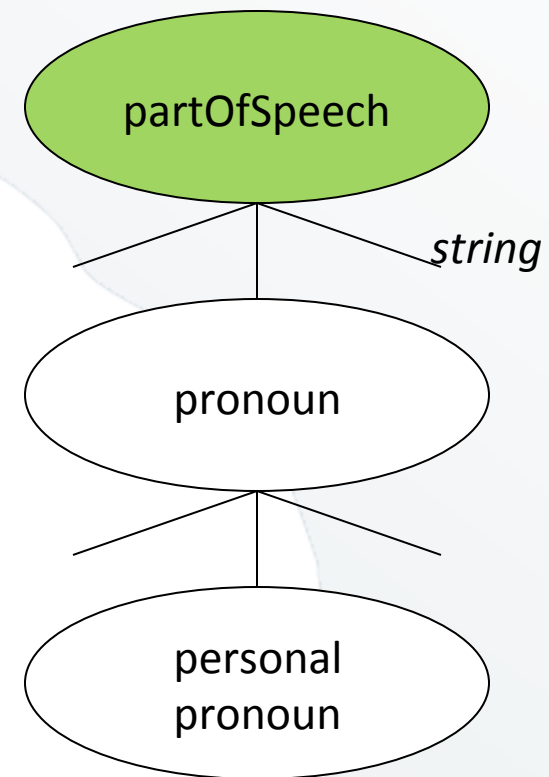
Data Category types

container:



Data Category relationships

- Value domain membership
- Subsumption relationships between simple data categories (legacy)
- Relationships between complex/container data categories are not stored in the DCR



Data categories and linguistic resources

Language	BWO	gram	rs

wordOrder ●

grammaticalGender ●

A (schema for a) typological database

writtenForm ●

Lemma

writtenForm ●

grammaticalGender ●

lexicalType ●

Word Form

Word Form



Lexicon

1..*

Lexical Entry

partOfSpeech ●

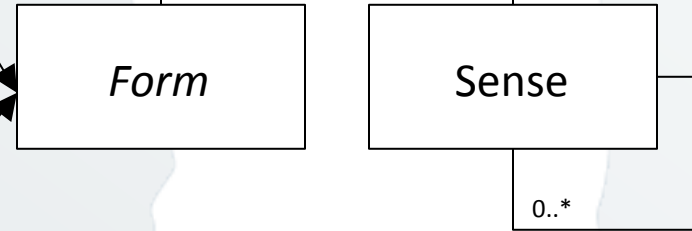
1..*

Form

0..*

Sense

0..*



A (schema for a) lexicon

The role of data categories in CMDI

- CMD components, elements and items can have links to *concepts*
- These links should be *resolvable* to a concept description
- This concept description gives *explicit semantics*
- Elements and components can use different terminology but still have *common semantics*

- ISOCAT provides resolvable links to the semantic description of data categories (DCs)
 - CMD items: simple DCs
 - CMD elements: complex DCs
 - CMD components: container DCs (upcoming in the Component Registry)

Data category references in CMDI

```
<CMD_Component name="HeadWordType"  
  ConceptLink="...">  
  <CMD_Element name="HeadWordType" ConceptLink="  
    http://www.isocat.org/datcat/DC-2486">  
    <ValueScheme>  
      <enumeration>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-286">Lemma</item>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-2948">Word form</item>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-350">Phrase</item>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-1386">Sentence</item>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-2599">Other</item>  
        <item ConceptLink="http://www.isocat.org/datcat/DC-2592">Unspecified</item>  
      </enumeration>  
    </ValueScheme>  
  </CMD_Element>  
</CMD_Component>
```

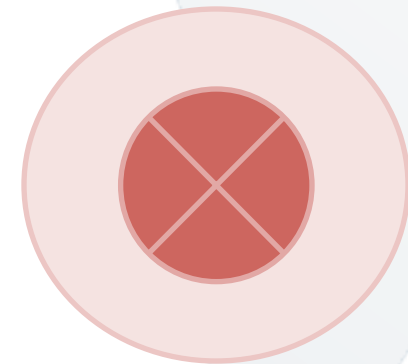
A glimpse of ISOcat



<http://www.isocat.org/>

Data Category Registry

- A (coherent) set of Data Categories, in our case for linguistic resources
- A system to manage this set:
 - Create and edit Data Categories
 - Share Data Categories, e.g., resolve PID references
 - Standardize Data Categories
- Grass roots approach



Thematic Domain Groups

TDG 1: Metadata

TDG 2: Morphosyntax

TDG 3: Semantic Content Representation

TDG 4: Syntax

TDG 5: Machine Readable Dictionary

TDG 6: Language Resource Ontology

TDG 7: Lexicography

TDG 8: Language Codes

TDG 9: Terminology

TDG 11: Multilingual Information Management

TDG 12: Lexical Resources

TDG 13: Lexical Semantics

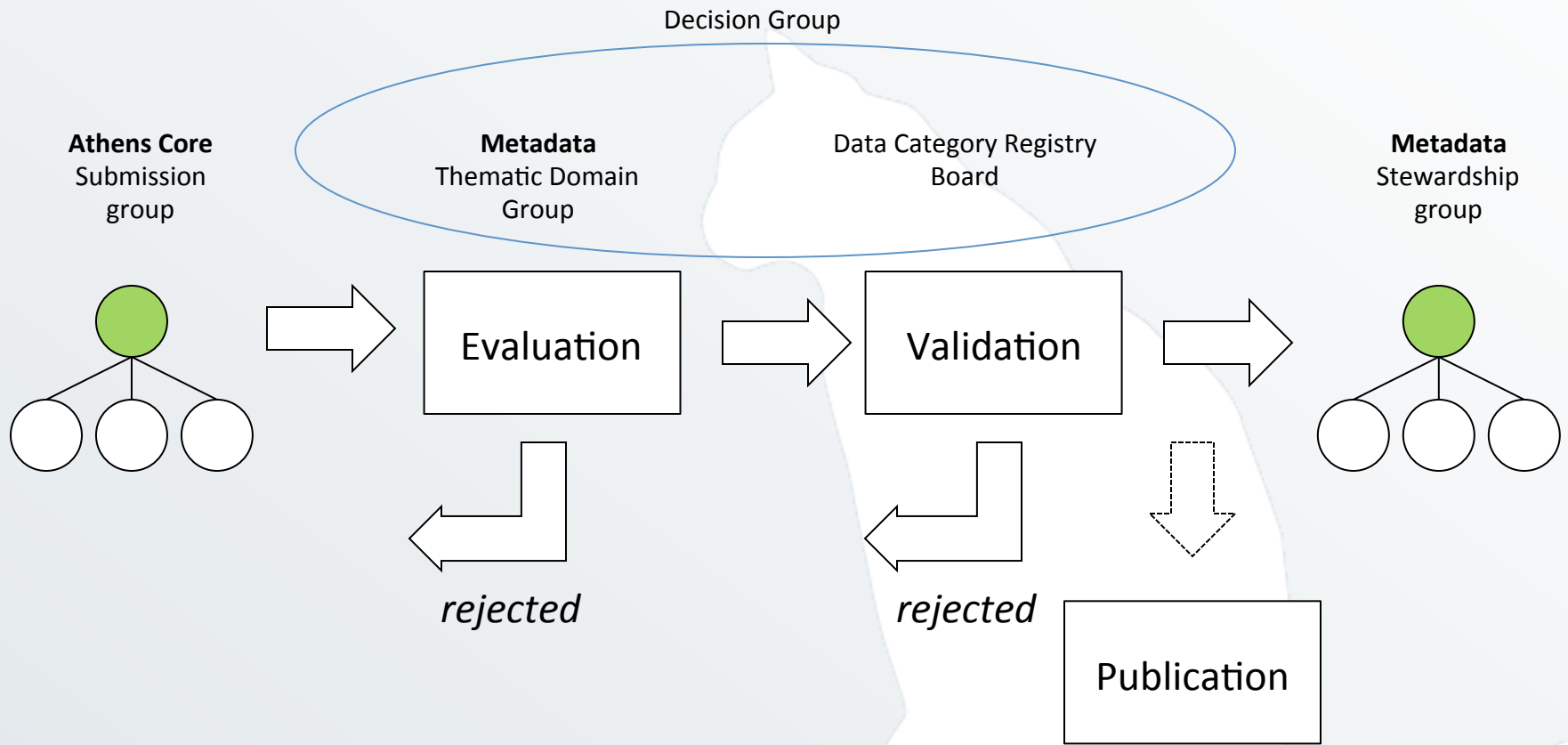
TDG 14: Source Identification

- TDGs are the caretakers of a (coherent) subset of the DCR
- TDGs own one or more profiles
- Each TDG has a chair
- A number of judges (assigned by SC P members)
- A number of expert members (up to 50%)
- TDGs are constituted at the TC37/SC plenary
- New TDGs need to be proposed by a SC
 1. Translation
 2. Sign language
 3. Audio

Status of the metadata profile

- Initial set of data categories has been created (to never disappear) by the Athens Core group
 - But your own components might need your own specific DCs
- Translations for many EU languages have been added
- No ISO Standardization yet ☹
 - In the ISOCat sandbox:
 - Athens Core group has submitted the first set of DCs
 - Metadata TDG is starting up the standardization process
- The addition of container DCs to be linked to CMD components is planned

Standardization



Component Registry

- Interacts with ISOCat
 - Access to *public* DCs in the metadata profile
 - So to currently access your private DCs you'll have to make them public
- Working towards:
 - The ability to access your private workspace from the Component Registry
 - Needs support for delegation of the users authentication, which can't be done with the current Shibboleth setup
 - Still will have to make your DCs public if you make your component/profile public

Thank you for your attention!

Visit

www.isocat.org

Questions?

isocat@mpi.nl

or

The CLARIN-D ISOcat tutorial end of 2011