ISOcat
www.isocat.org

ISOcat introduction

ISOcat: a Data Category Registry

- An 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
    - Successor to ISO 12620:1999 which contained a hardcoded list of Data Categories

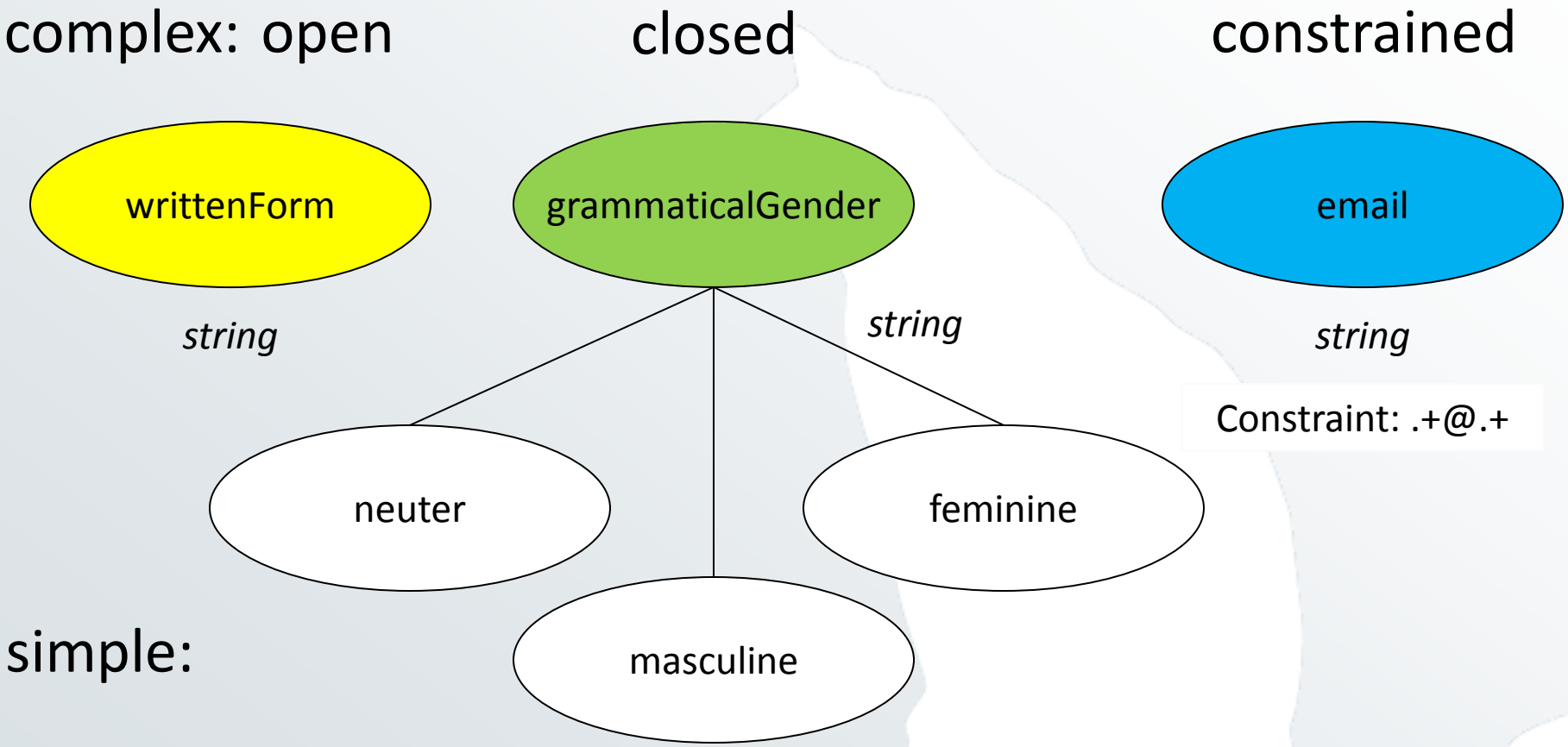
- A data category
  - is the result of the specification of a given data field
  - an elementary descriptor in a linguistic structure or an annotation scheme

Data Category example

- Data category: */grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
  - Descriptive part:
    - English name: grammatical gender
    - 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.
  - Linguistic part:
    - Conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

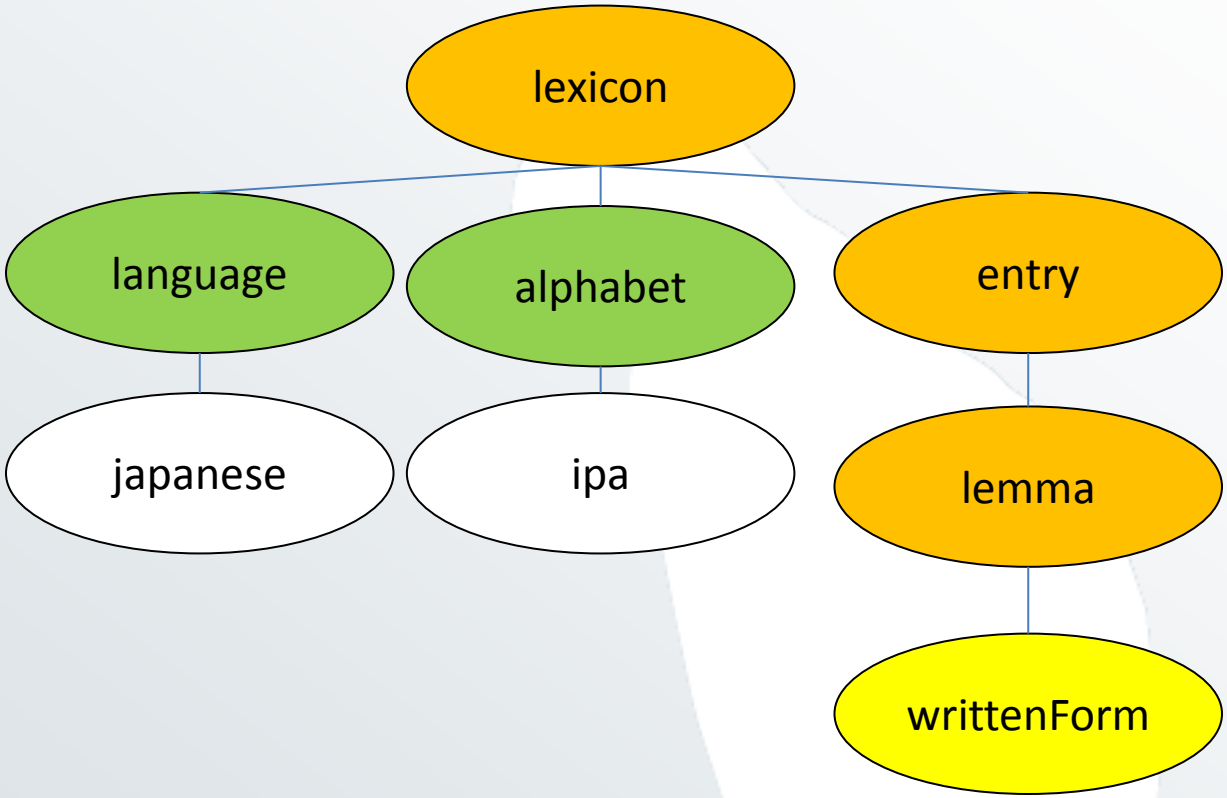
ISOcat
www.isocat.org

Data Category types



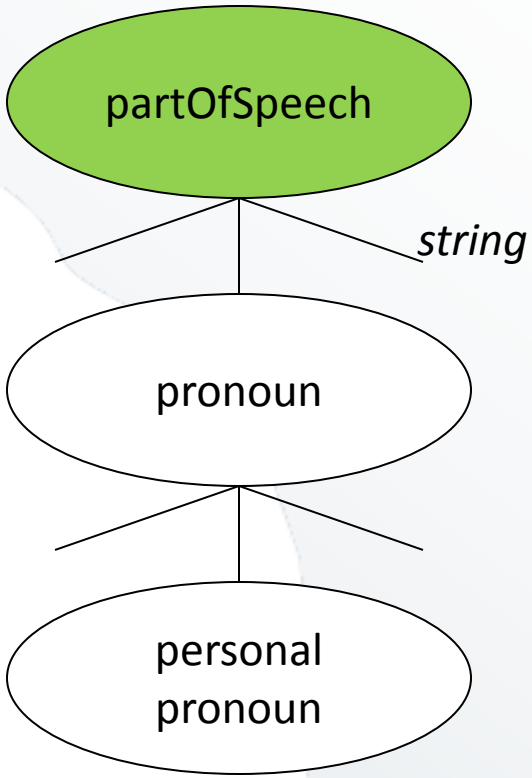
Data Category types

container:

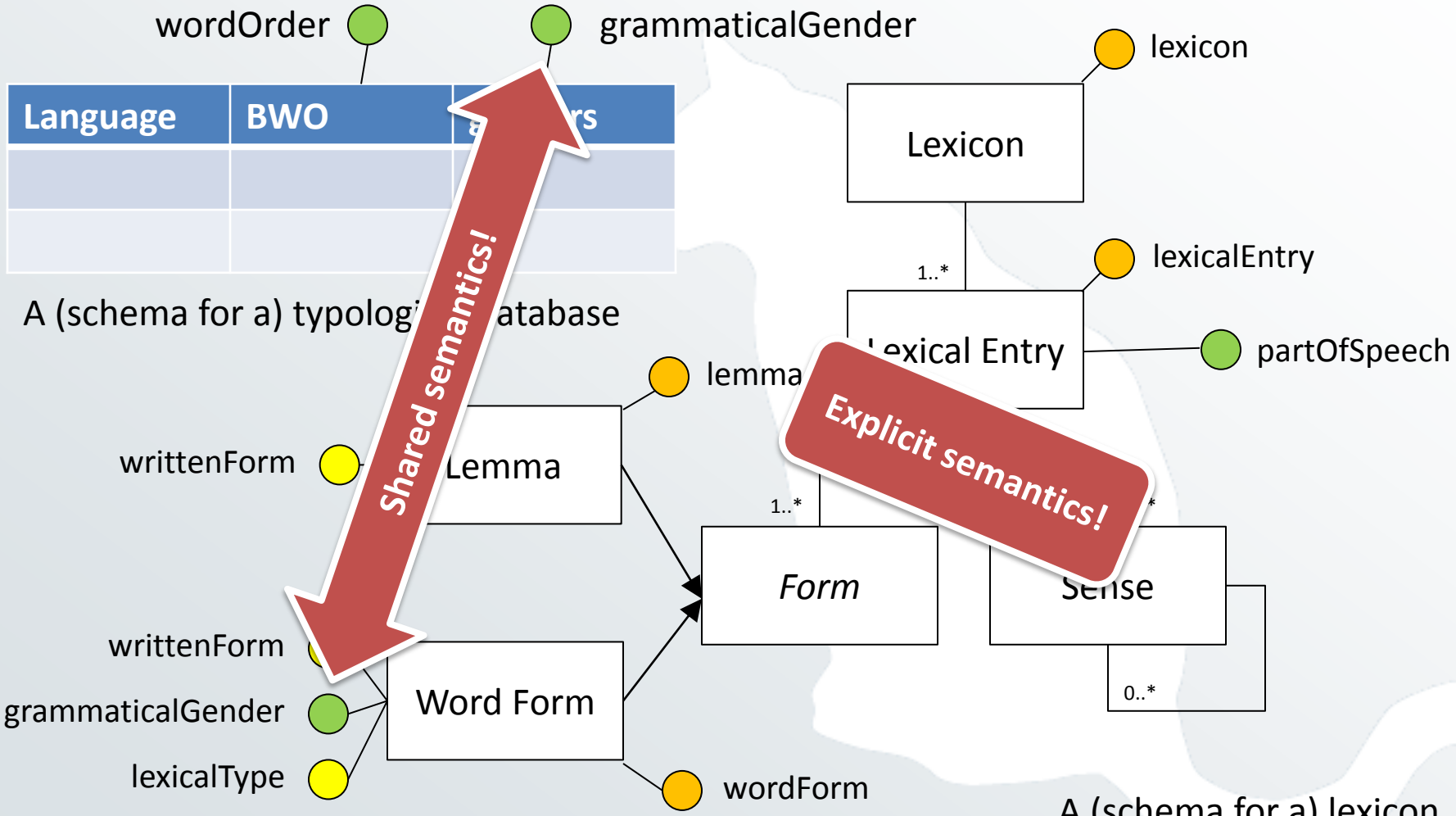


Data Category relationships

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



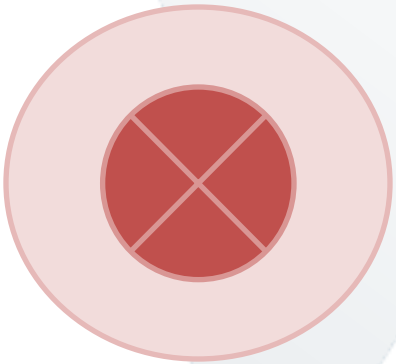
Data Categories and semantics



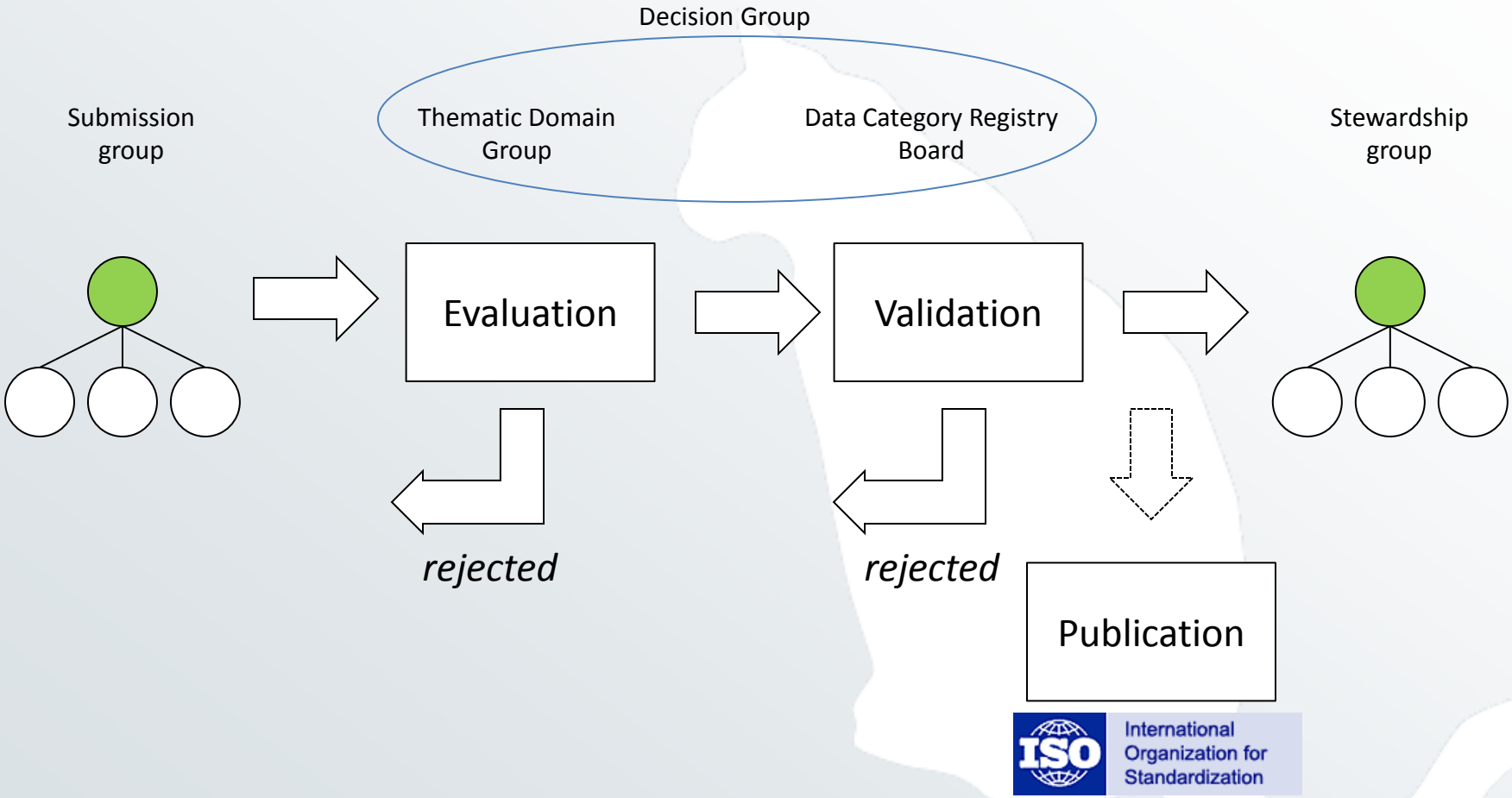
ISOcat
www.isocat.org

A 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



Standardization



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 owner and guardians 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

Status of standardization

- Unfortunately the standardization process hasn't taken off fully
  - there are still no standardized ISOcat data categories
- Groups in ISOcat will get more functionality
  - group specific views only show categories selected by the group
    - http://www.isocat.org/interface/index.html?view=CLARIN-NL/VL
  - 'used by'/'liked by'/'recommended by' can give an indication of reuse/popularity
  - …

How can you use a Data Category Registry?

- You can:
  - Find Data Categories relevant for your resources and embed references to them so the semantics of (parts of) your resources are made explicit
    - This can be supported by tools you use, e.g., ELAN, LEXUS and the CMDI Component Editor directly interact with ISOcat
  - Interact with Data Category owners to improve (the coverage of) their Data Categories
  - Create (together with others) new Data Categories and/or selections needed for your resources and share those
  - (Submit (your) Data Categories for standardization)
  - Make the semantics of your resources explicit and shared
  - Free of charge
  - Grass roots approach