www.isocat.org

ISOcat introduction

www.isocat.org

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

What is a 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.*

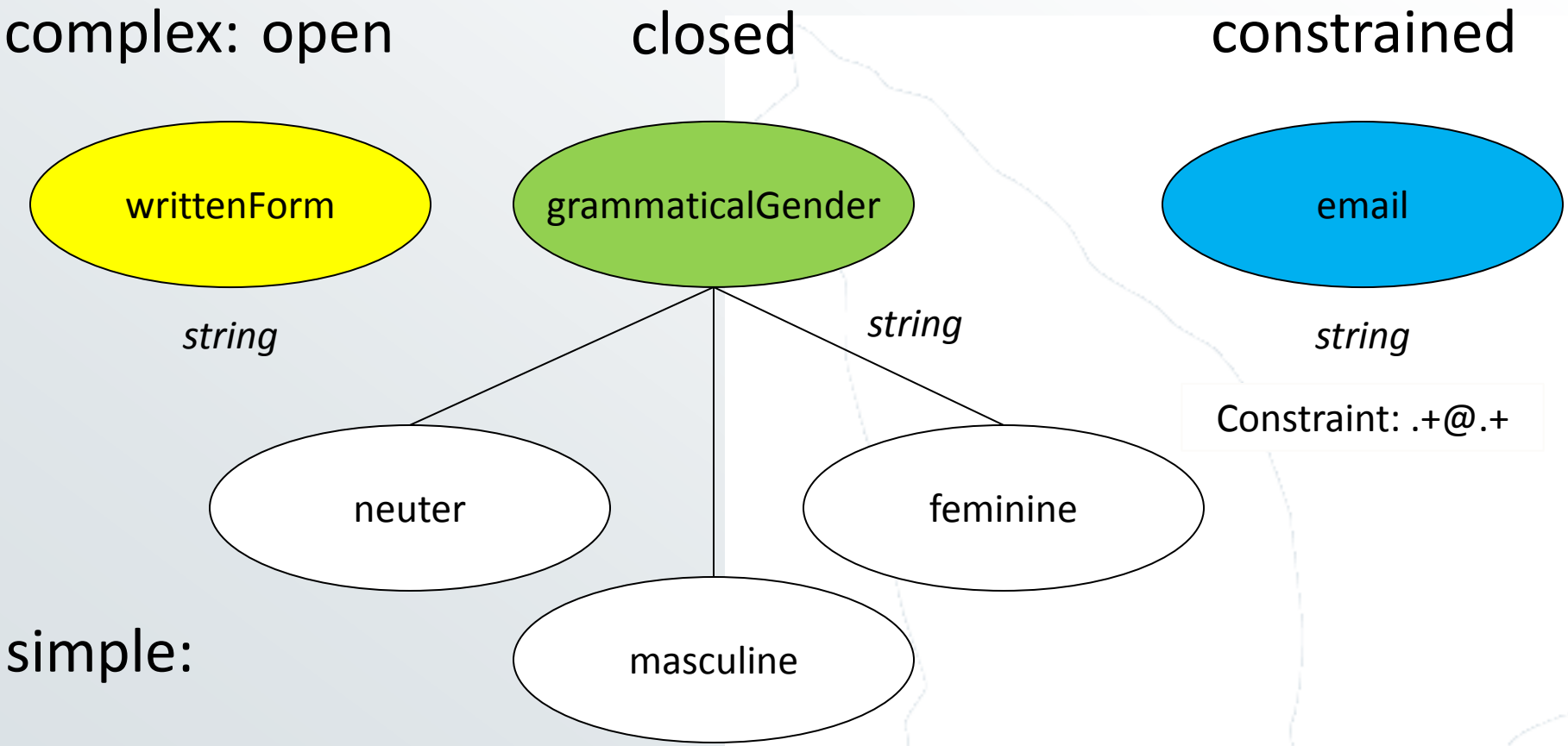
- Specification consists of 3 main parts:
  - *Administrative part*
    - *Administration and identification*
  - *Descriptive part*
    - *Documentation in various working languages*
  - *Linguistic part*
    - *Conceptual domain(s for various object languages)*

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.
  - Linguistic part:
    - Morposyntax conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

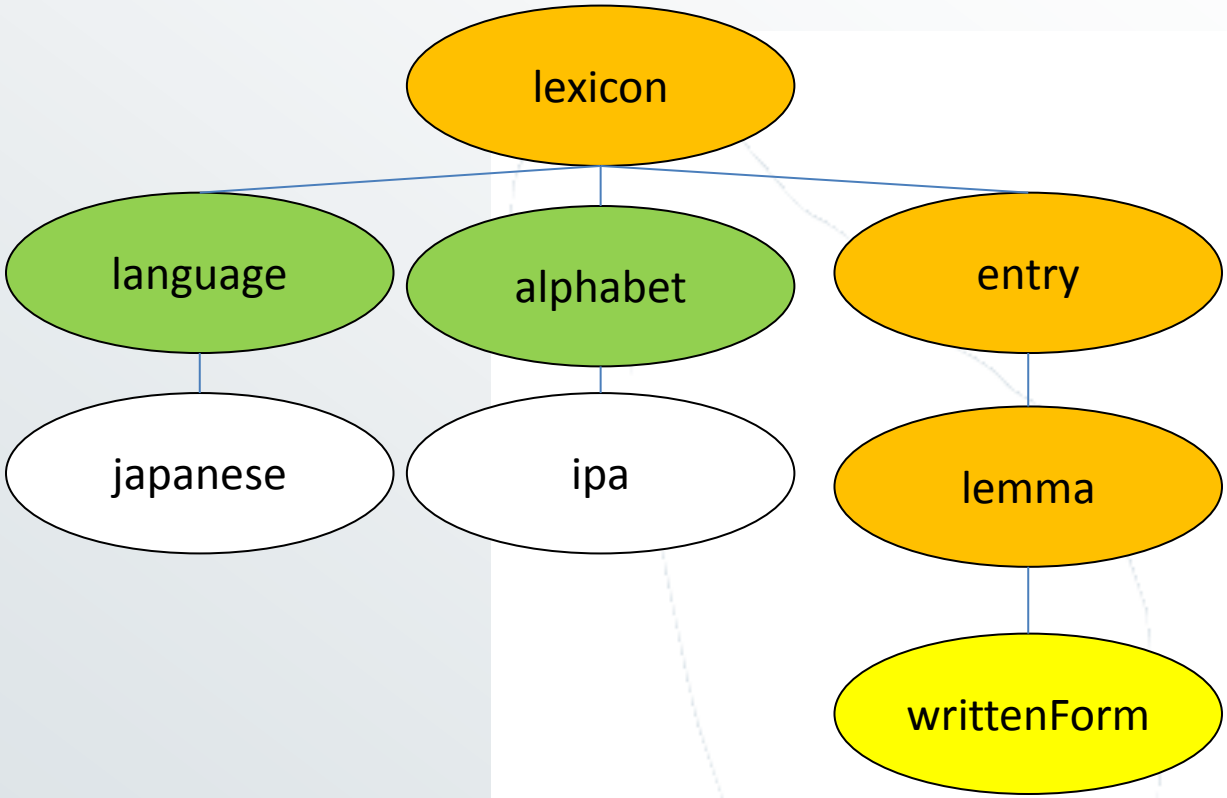
www.isocat.org

Data Category types



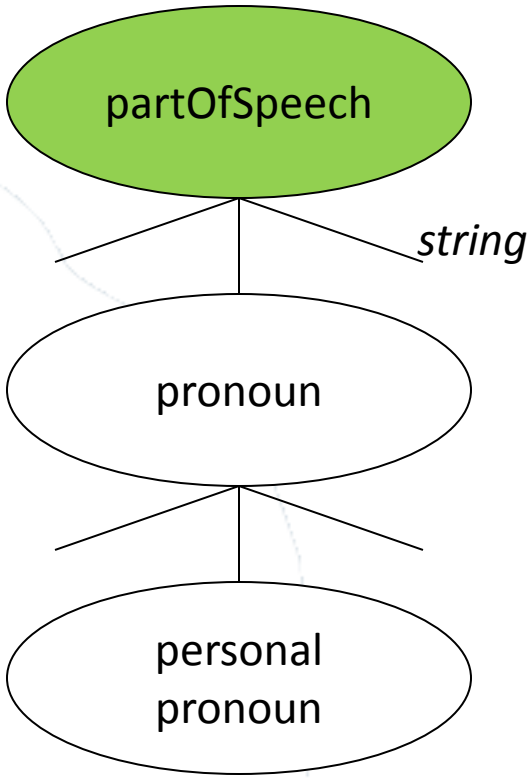
Data Category types

container:



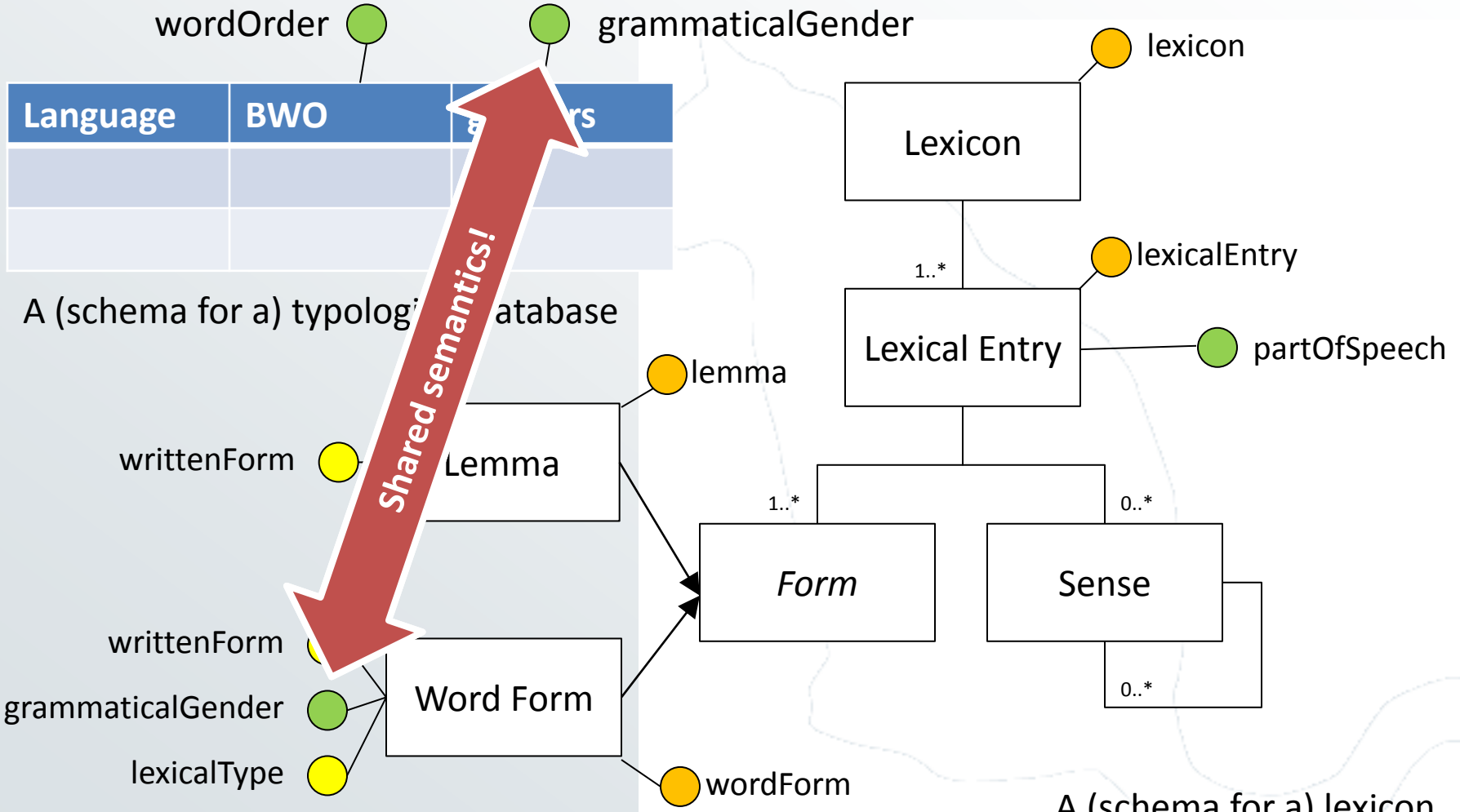
Data Category relationships

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



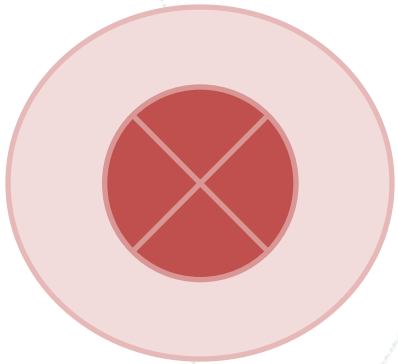
www.isocat.org

How can you use Data Categories?

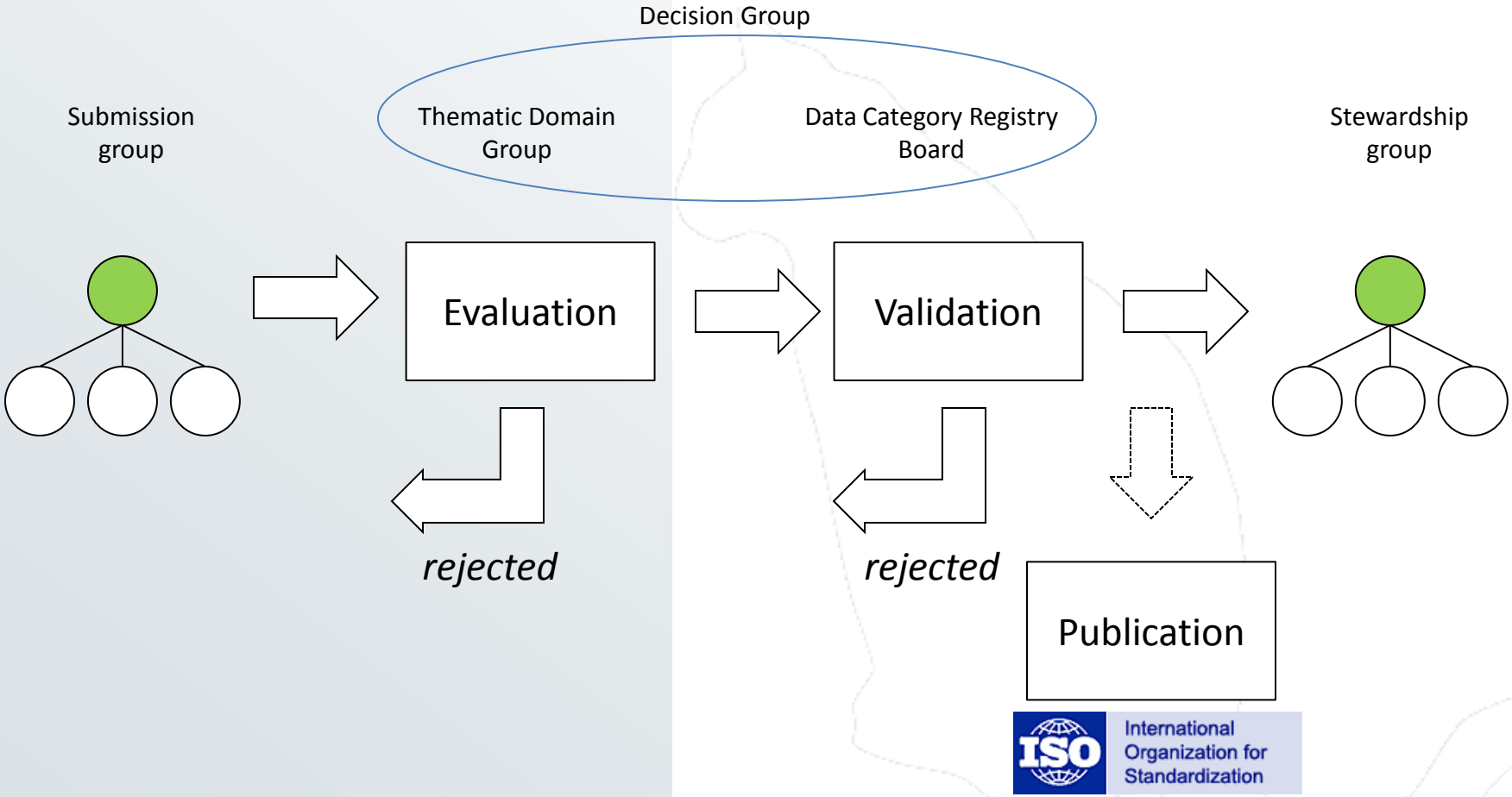


What is a Data Category Registry?

- 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



www.isocat.org

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
  3. Audio

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
  - Free of charge
  - Grass roots approach

ISOcat and CLARIN(-NL): general remarks

ISOcat
www.isocat.org

Importance of ISOcat

- Collaboration
  - Human, machine, language x, language y

Essential in CLARIN, but …

Impossible when we don't know (exactly) what we are talking about!

- Transitive verb – transitief werkwoord
- Transitief werkwoord – overgankelijk werkwoord

ISOcat
www.isocat.org

Importance of ISOcat

- ISOcat:
  - Provides us with a framework to make such things clear (*is X the same as Y, does A use it the same way*)
  - At least, that is the intention, ISOcat still being 'under construction'
- Today's sessions:
  - How to work with ISOcat
  - Which other "cats" do we have at the moment
  - The future …

ISOcat
www.isocat.org

CLARIN-NL (and VL) and ISOcat

- There are some 25 projects dealing with ISOcat in some sense (sometimes ‘only’ metadata)
  - 20 Netherlands
  - 3 Flanders
  - 1 NL/VL pilot
  - Of course, that is not the main focus of these projects, but still…
  - A lot of ISOcat work needs to be done!

ISOcat
www.isocat.org

CLARIN-NL (and VL) and ISOcat

- At least of TTNWW (the pilot) one of the explicit goals is to signal problems and to try to remedy them (for our own good, and that of CLARIN as a whole)

- In that respect, we do have some ‘success’
  - Several larger and smaller issues are already being remedied

ISOcat
www.isocat.org

CLARIN-NL (and VL) and ISOcat

Many (Dutch) projects working on ISOcat issues, plus those of other national CLARINs

- same concepts ?
- same problems ?

⇒very likely

ISOcat
www.isocat.org

Collaboration necessary

- National (Dutch) level
  - Coordinated effort
  - Shared workspace under ‘shared’
    - **USE IT**

    Plus discussion platform
  - Report problems to me (Ineke)
- International level
  - We will try to collaborate with them as well

ISOcat
www.isocat.org

Thanks !