Metadata and the ISO DCR

PeWi based on the stuff from

Marc Kemps-Snijders, Sue Ellen Wright, Menzo Windhouwer

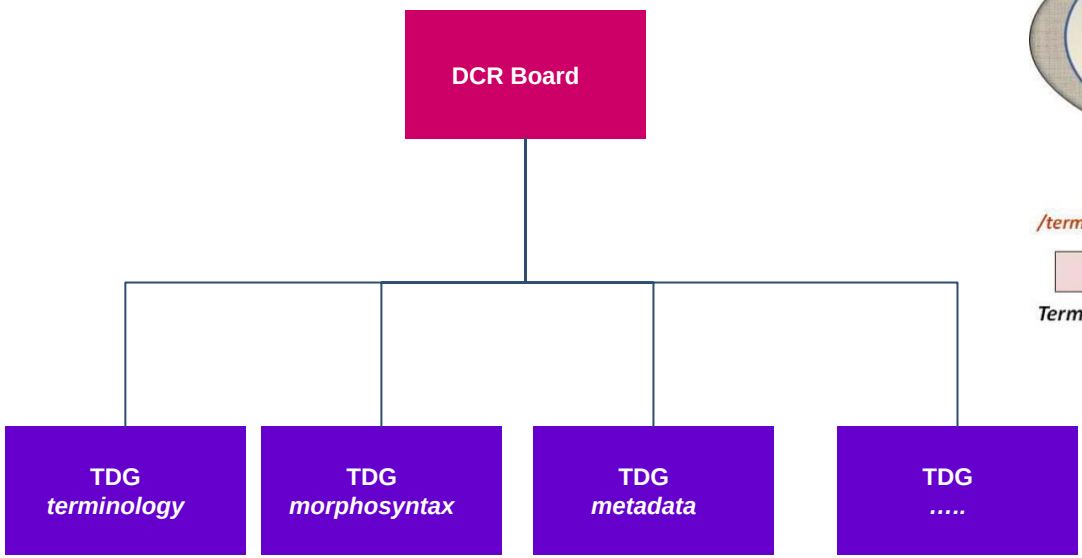
Athens
2009-01-30/31

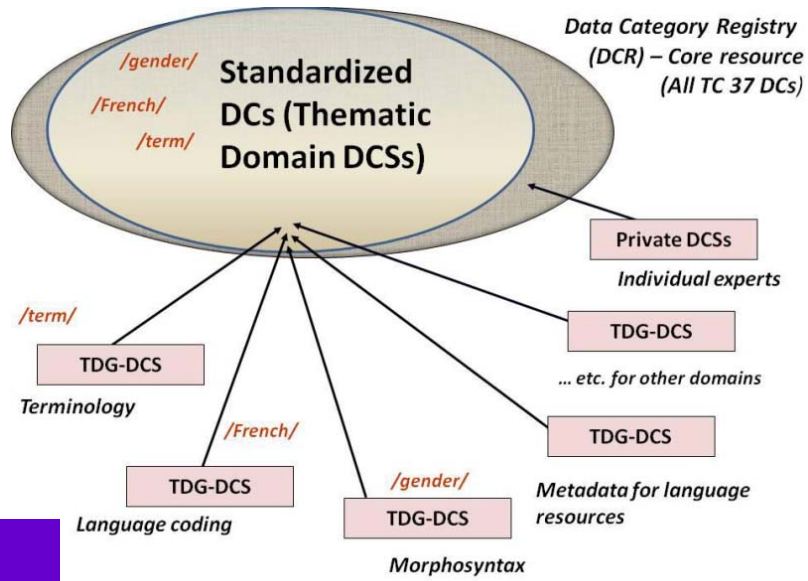
General Stuff

- ISOcat is the reference implementation of ISO 12620:2009
- ISO 12620 is in ?? state - so fairly much stable
- Will use ISOcat within CLARIN as far going as possible knowing that it will be a long time to sort out all aspects
- will use ISOcat for metadata, i.e. all relevant concepts nit registered elsewhere should be in ISOcat
- see this expert group as a first step to a board governing the Metadata Profile in ISOcat
- Pewi is Thematic Domain Group leader ☺
- of course we need to ask the Asians, the US etc

DCR Structure and Process

- 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.

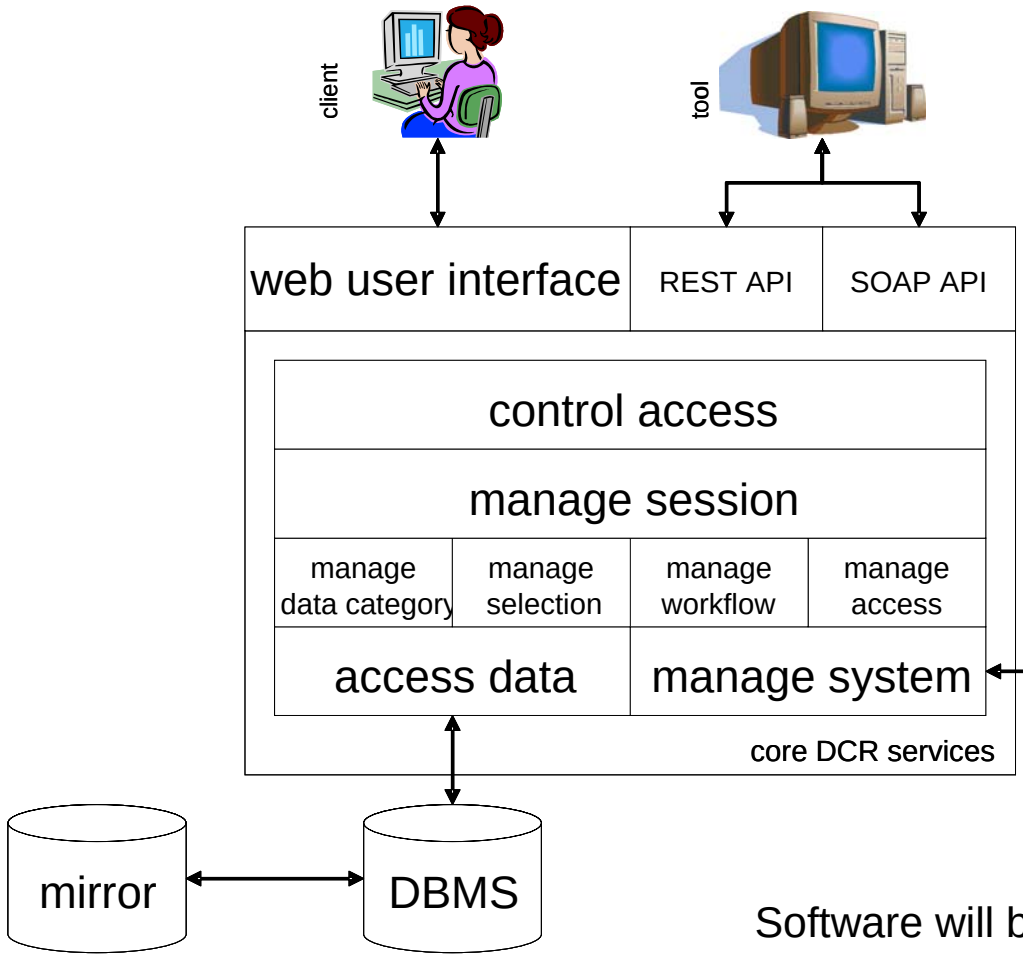
ISOcat

TDG Structure

TDG 1: Metadata
TDG 2: Morphosyntax
TDG 3: Semantic Content Representation
Activity 1: Discourse Relations
Activity 2: Dialogue Acts
Activity 3: Referential Structures and Links
Activity 4: Logico-semantic Relations
Activity 5: Temporal Entities and Relations
Activity 6: Semantic Roles and Argument Structures
TDG 4: Syntax
TDG 5: Machine Readable Dictionary
TDG 6: Language Resource Ontology
TDG 7: Lexicography
TDG 8: Language Codes
TDG 9: Terminology
Activity 1: General Principles
Activity 2: Concept Modeling
Activity 3: ISO Terminology Entries
Activity 4: Benchmarking Terminology
Activity 5: Terminology Management
Activity 6: TBX
Activity 7: TBX-Basic
Activity 8: Other TBX/TMLs
Activity 9: Geneter
Activity 10: TMS
TDG 11: Multilingual Information Management
TDG 12: Lexical Resources
TDG 13: Lexical Semantics
TDG 14: Source Identification

ISOcat

ISOcat Architecture



Referencing:
http://www.isocat.org/datcat/DC-1708

DCS Export:
Basic RDF (implemented), Relax NG (planned), XML Schema (planned), OWL (planned), XCS (planned), ODD (planned)

Software will be open, i.e. anyone can setup a concept registry

ISOcat

Data category

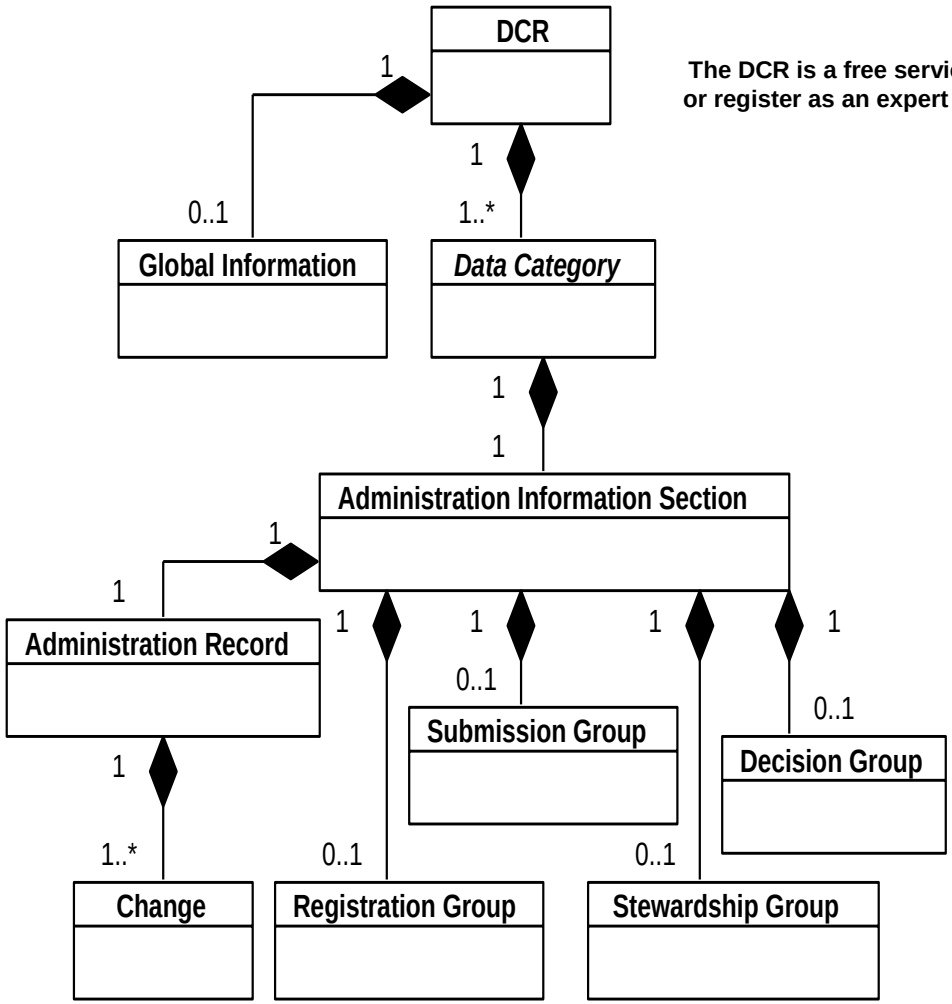
- The result of the specification of a given data field
  - *A data category is an elementary descriptor in a linguistic structure or an annotation scheme.*

- Model consists of 3 main parts:
  - *Administrative part*
    - *Administration and identification*
  - *Descriptive part*
    - *Documentation and working language*
  - *Linguistic part*
    - *Conceptual domain of object language*

ISOcat

Data category

Administrative part

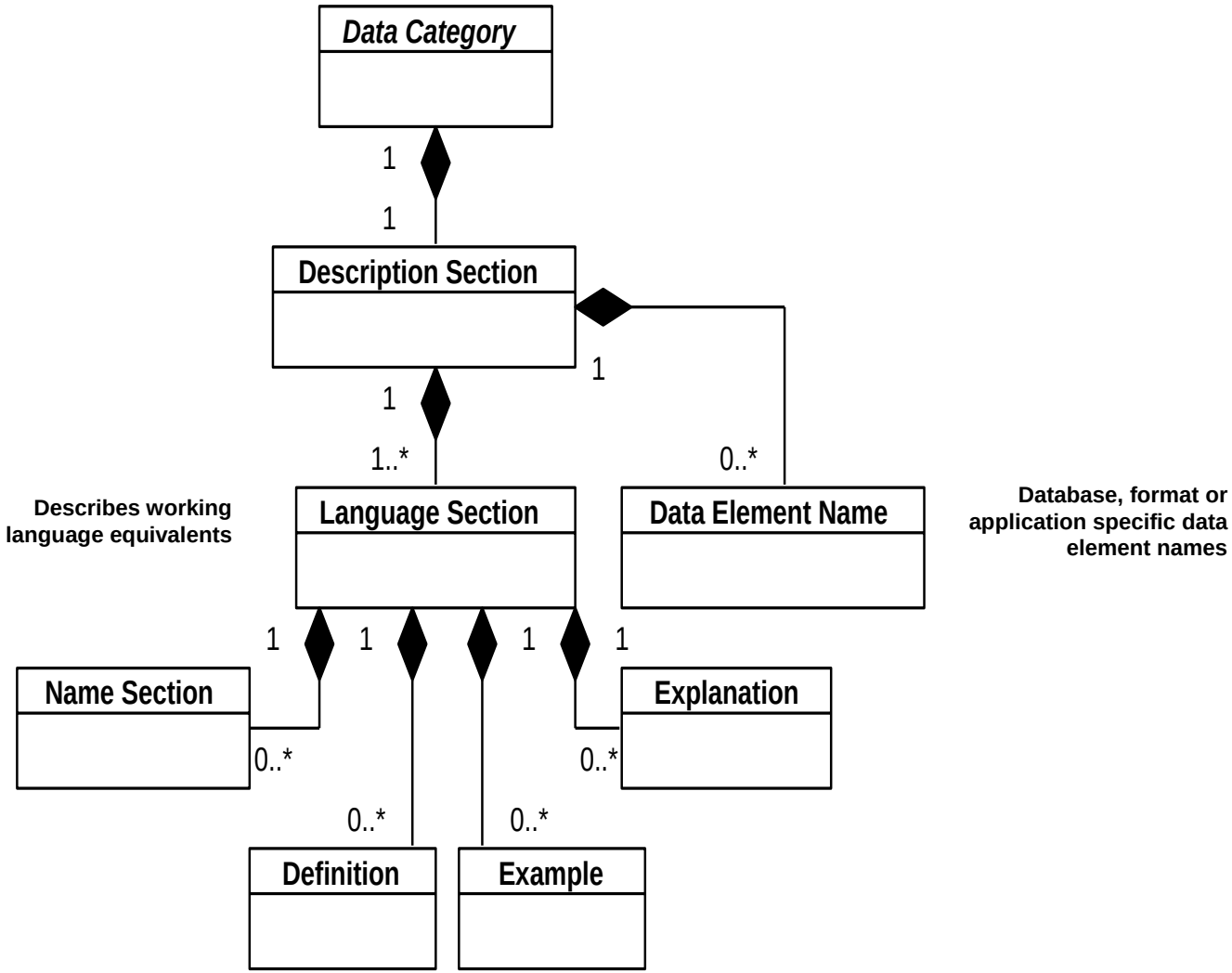


The DCR is a free service: anyone can access it or register as an expert and create/share his/her own data categories.

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

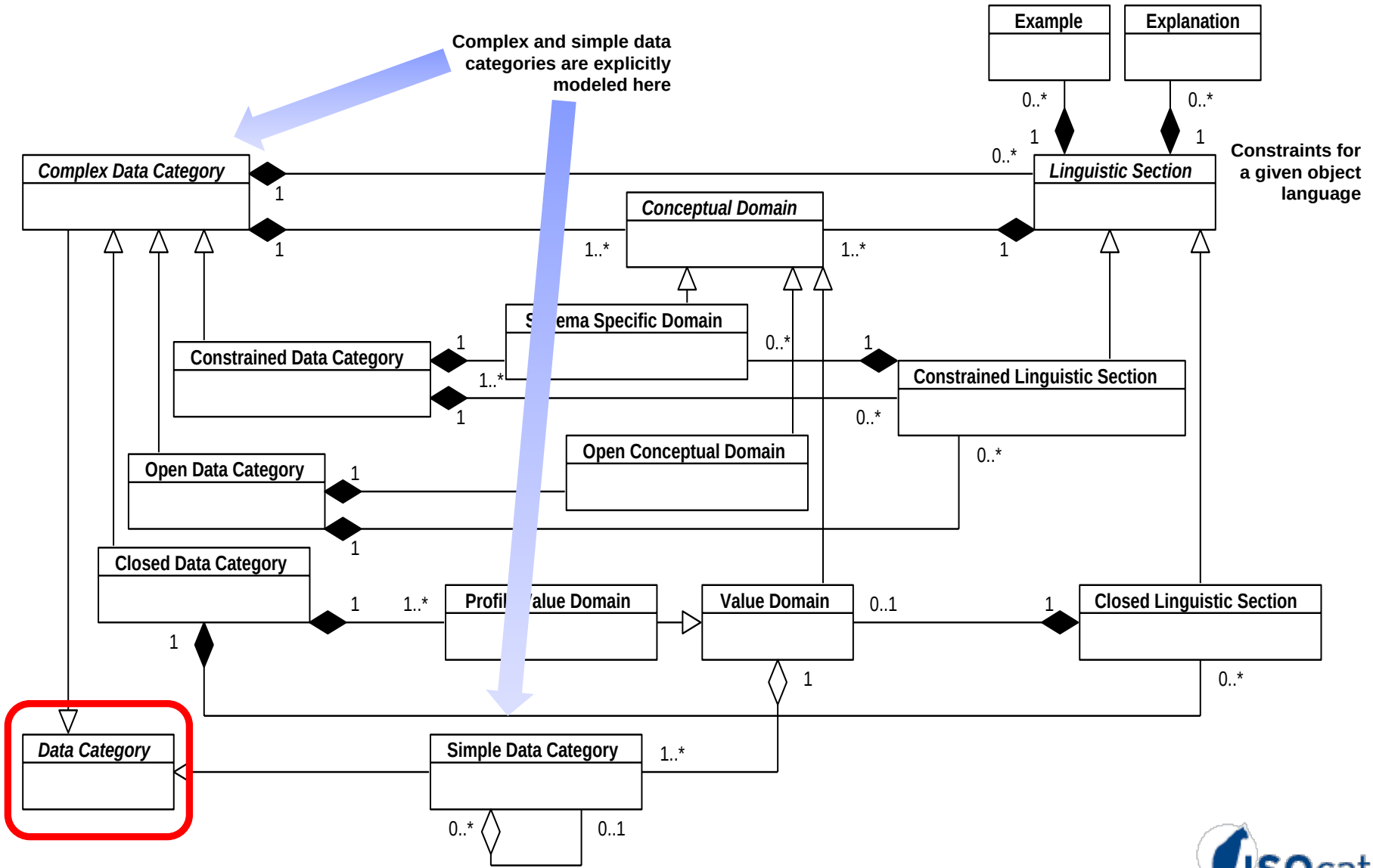
Data category

Descriptive part



Data category

Linguistic part



Data category

Linguistic part (example)

- Data category: */Continent/*
  - Conceptual domain: */Africa/*, */Europe/*, */South America/ etc*
    - *Lists all agreed values*
  - Linguistic Section
    - Language: ger
    - Value Domain: */Afrika/*, */Europa/*, */Süd Amerika/ etc*
    - *Lists all agreed values for German*

Known Problems - granularity

- semantic granularity - old issue: compare with DC

  15 general elements vs. qualified elements
- so for us:
  - is */date/* sufficient or do we need */date of birth/*, */date of creation/*, */date of publication/* etc
  - is */language/* sufficient or do we need */language of speaker/*, */language a document is in/*, */language a document is about/*, etc
  - or can we rely on the contextual embedding, i.e.

    /person.language/, /person.date/, /content.language/, etc
- task for us: define all the elements that are required to describe a certain linguistic resource type
- there needs to be some knowledge engineering

ISOcat

Known Problems - relations

- ISO 12620 does not allow to specify relations except for more generic concepts so */transitive verb/ is_a /verb/*
- need a framework that allows users to easily specify

  "for me now the following holds:

  DC:creatorname = IMDI:actorname"

ISOcat

Let's get it working now ☺