ISOcat
www.isocat.org

# Data Category specifications

# Mandatory parts of the specification to be provided by the user

- For each data category:
  - a mnemonic identifier
  - an English definition
  - an English name
- For complex data categories:
  - a conceptual domain
- For standardization candidates:
  - a profile (other then Private)
  - a justification

# Data Category example

- Closed data category: */grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
    - Justification: Used in Morphosyntax, Terminology, Lexicography
  - Descriptive part:
    - Profile: Morphosyntax
    - English name: grammatical gender
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
  - Linguistic part:
    - Morphosyntax conceptual domain: */male/, /feminine/, /neuter/*

# Data category identifiers

- a mnemonic string used to refer to the data category
  - not unique (PIDs are unique), i.e., multiple categories with the same identifier might exist due to different owners, thematic domains, versions, ....
- should be based on a meaningful English word or series of words presented as an alphanumeric character string; for multiword strings, begin with lowercase and express the identifier as one continuous string in camel case with no white space
- maybe used in XML vocabularies and thus must be a valid local part of a qualified name:
  - cannot start with a number, shouldn’t contain any whitespace, …
- for example: not /*1stPerson*/ but /*firstPerson*/, not /*EVON*/ but /*singularNeuterForm*/

# Persistent Identifiers

- Each Data Category should be uniquely identifiable
  - Ambiguity: different domains use the same term but mean different ‘things’
  - Semantic rot: even in the same domain the meaning of a term changes over time
  - Persistence: for archived resources Data Category references should still be resolvable and point to the specification as it was at/close to time of creation
- Persistent IDentifier
  - ISO 24619:2011 Language resource management -- Persistent identification and access in language technology applications
  - ISOcat uses ‘cool URIs’
    - http://www.isocat.org/datcat/DC-1297 (/*grammaticalGender/*)
  - managed by the system

# Justification

- a short description justifying why the data category should be included in the registry
- mandatory for data categories to be standardized; desirable in general
- even data categories that are common in a given thematic domain may be unfamiliar or ambiguous to users unfamiliar with that domain

# Thematic Domain Groups/profiles

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

# Data Element Name Sections

- used to record names for the data category as used in a given database, format or application
- language independent
- attributes:
  - the mandatory data element name
    - one identifier (word, multi-word unit or (alpha)numeric representation
  - the mandatory source
    - the database, format or application in which the data element name is used
- proper place to mention abbreviations/tags used for a particular notion, and not just for English: N, NPlur, EVON

# Data Category example

- Closed data category: */grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
    - Justification: Used in Morphosyntax, Terminology, Lexicography
  - Descriptive part:
    - Profile: Morphosyntax
    - Data Element Name: GramGender in Text Meaning Representation
    - English name: grammatical gender
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
  - Linguistic part:
    - Morphosyntax conceptual domain: */male/, /feminine/, /neuter/*

# Working and object languages

- Working language:
  - language used to describe objects
- Object language:
  - language being described

You can describe properties of the object language French in the working language Dutch:

*In de Franse taal worden vrouwelijke en mannelijk zelfstandige naamwoorden onderscheiden.*

# Language sections

- Provide the correct full name(s) in the working language at hand
  - mark the status of the name: deprecated, admitted, preferred, standardized
- Provide one precise definition in the working language at hand
  - the data model provides place for multiple definitions, but this just leads to confusion

# Data Category example

- Closed data category: */grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
    - Justification: Used in Morphosyntax, Terminology, Lexicography
  - Descriptive part:
    - Profile: Morphosyntax
    - Data Element Name: GramGender in Text Meaning Representation
    - English name: grammatical gender
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
    - French name: genre grammatical
    - French definition: Catégorie fondée (selon la langue) sur la distinction naturelle entre les sexes ou d'autres critères formels.
  - Linguistic part:
    - Morphosyntax conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

# Conceptual domains

- The mandatory data type
  - the data type, as defined for W3C XML Schema, of this complex data category
  - the default data type is *string*
- Additionally:
  - closed data categories:
    - set of permissible values (simple data categories) for each profile
  - constrained data categories:
    - constraint specified in a supported rule language
      - e.g., an XML Schema regular expression or facet

# Profile value domains and (hierarchies of) simple data categories

| Data category | Morposyntax | Terminology |
|---|---|---|
| ● */partOfSpeech/* | X | X |
| ○ */adjective/* | X | X |
| ○ */ordinalAdjective/* | X | |
| ○ */participleAdjective/* | X | |
| ○ */qualifierAdjective/* | X | |
| ○ */adposition/* | X | X |
| ○ */circumposition/* | X | |
| ○ */preposition/* | X | |
| ○ */postposition/* | X | |

# Bulk import

- The ISOcat system administrator can import bulks of new Data Categories or updates

  http://www.isocat.org/forum/viewtopic.php?f=3&t=14