www.isocat.org

ISOcat

An ISO 12620:2009 Data Category Registry

Marc Kemps-Snijders[a], Menzo Windhouwer[a], Sue Ellen Wright[b]

[a]Max Planck Institute for Psycholinguistics, [b]Kent State University

marc.kemps-snijders@mpi.nl , menzo.windhouwer@mpi.nl, sellenwright@gmail.com

ISOcat
www.isocat.org

Outline

- ISO 12620:2009
  - What are Data Categories?
  - How can you use Data Categories?
  - What is a Data Category Registry?
  - How can you use a Data Category Registry?
- ISOcat
  - Demonstration/Tutorial
- Future work

ISOcat
www.isocat.org

ISO 12620:2009

- Terminology and other content and language resources — Specification of data categories and management of a Data Category Registry for language resources
  - An ISO TC 37/SC 3 standard (see [1])
  - Successor to ISO 12620:1999 which contained a hardcoded list of Data Categories

www.isocat.org

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)*

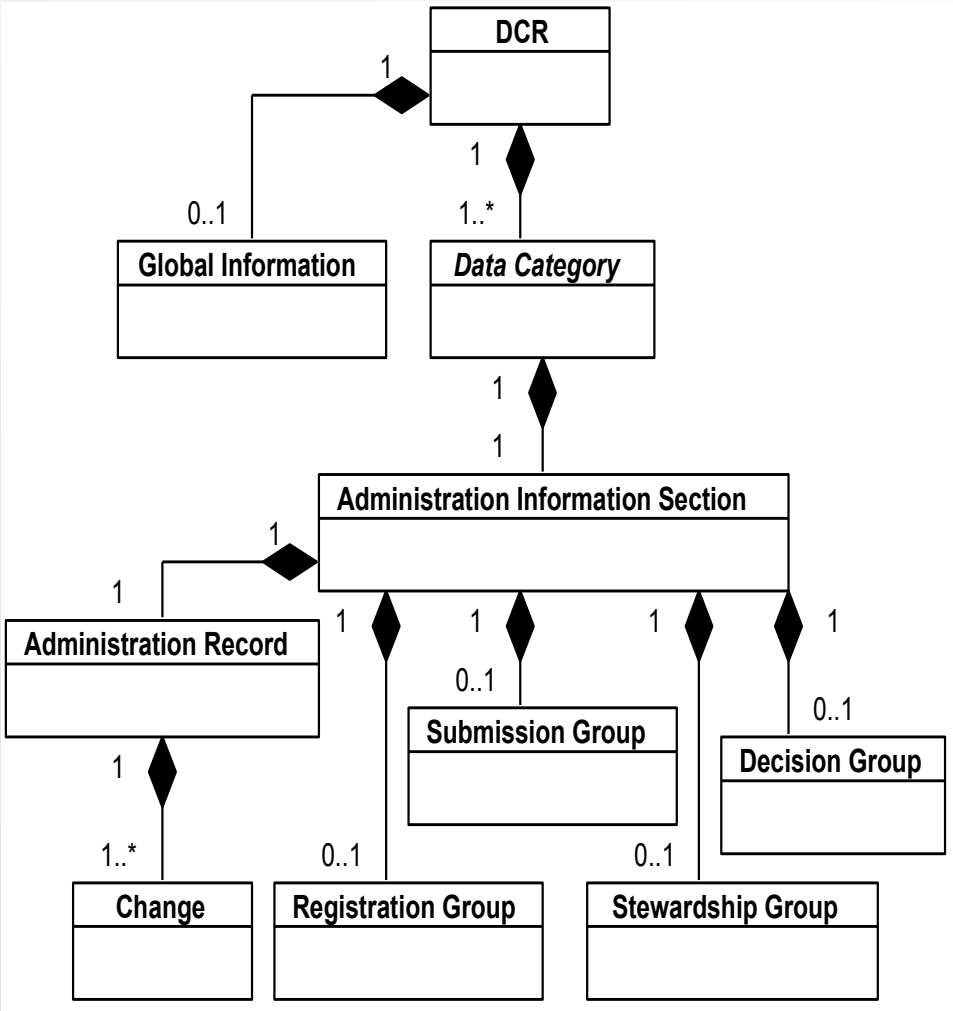
ISOcat
www.isocat.org

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/*

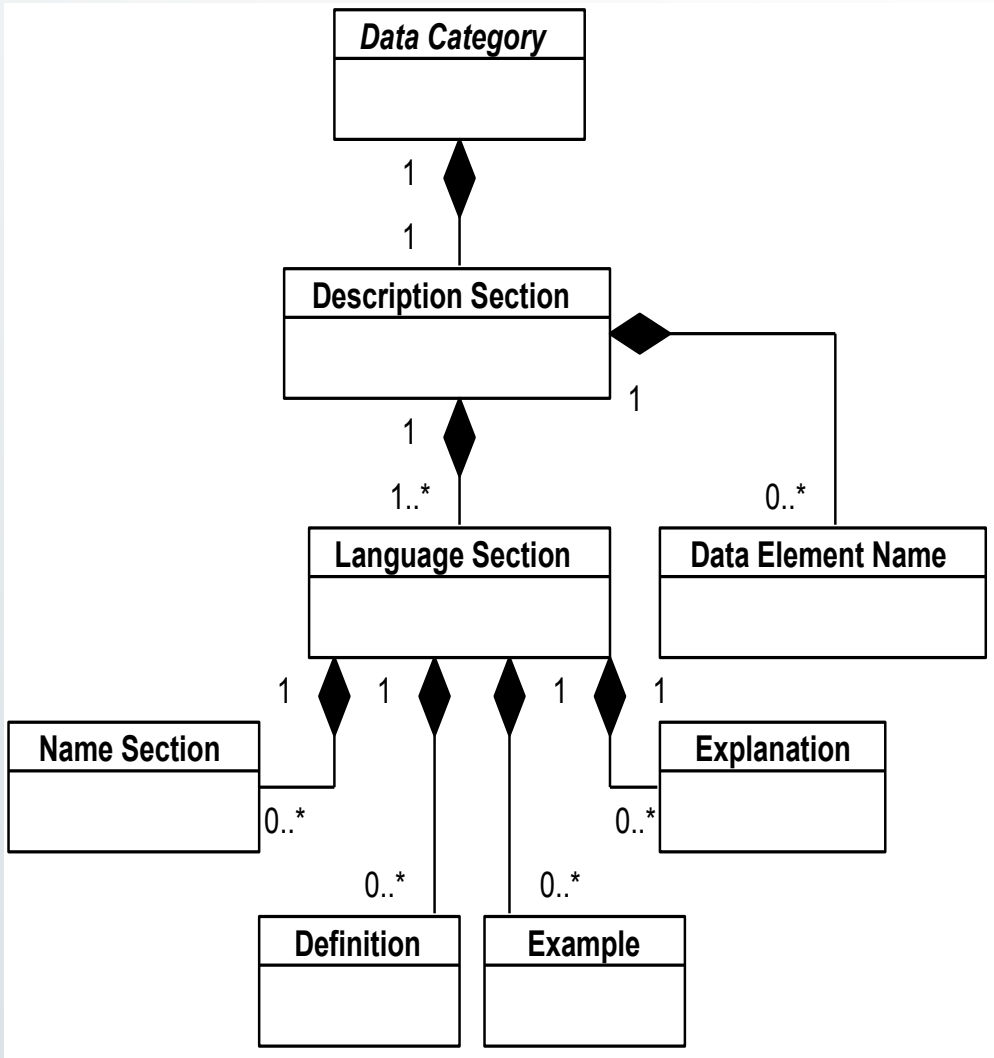
www.isocat.org

Data Category specification – Administrative part



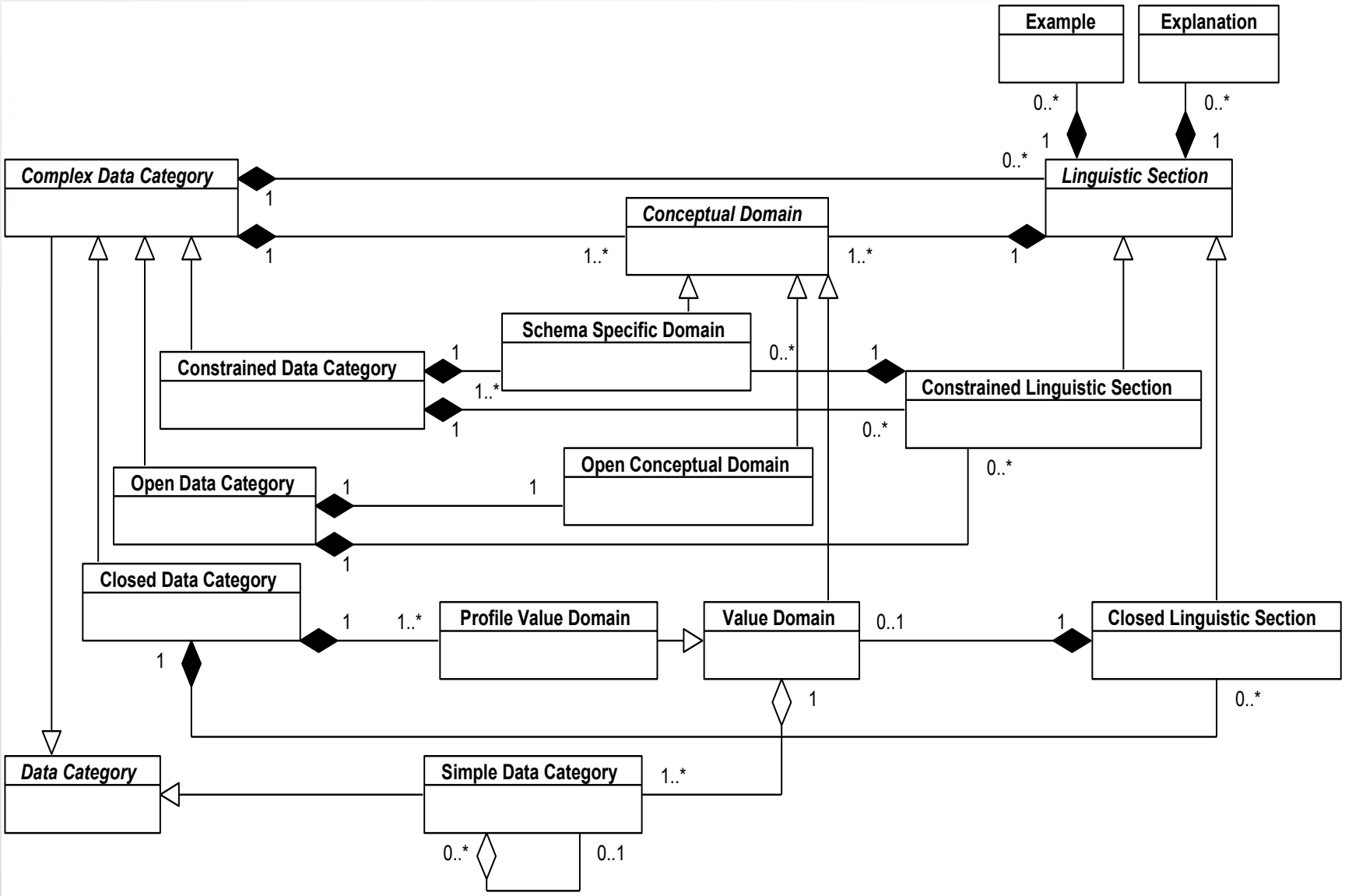
www.isocat.org

Data Category specification – Descriptive part



www.isocat.org

Data Category specification – Linguistic part



ISOcat
www.isocat.org

Mandatory parts of the specification

- 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
  - a justification

www.isocat.org

Guidelines for the specification

(see [2])

- Identifier:
  - camel case and XML-valid element name (without a namespace)
    - partOfSpeech
    - ~~my:POS~~, ~~123POS~~
- Data Element Name:
  - language independent name for the data category used in a specific application domain (specified in the source)
    - PoS in TBX
    - NN in myTagset or N in yourTagset (if widely used)

ISOcat
www.isocat.org

More guidelines

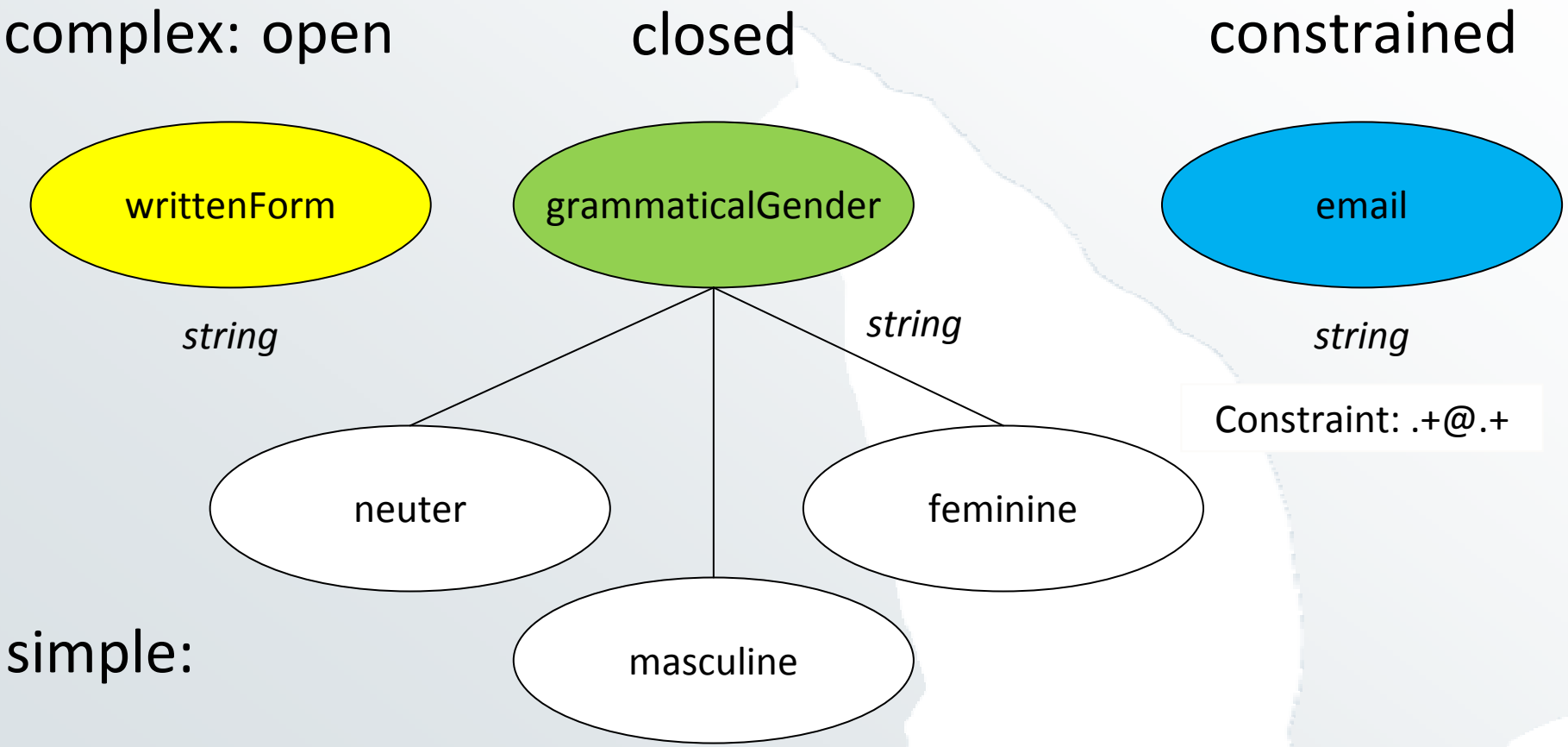
- Name Section in a Language Section
  - legible name
    - ‘part of speech’ in the English language section
    - ‘partie du discours’ in the French language section
- Definition:
  - intentional definitions (ISO 704)
  - should consist of a single sentence fragment
- Source:
  - add a source for any quoted material

More guidelines

- Justification:
  - a simple statement justifying the relevance of the data category to the field of language resources
  - especially needed for standardization

www.isocat.org

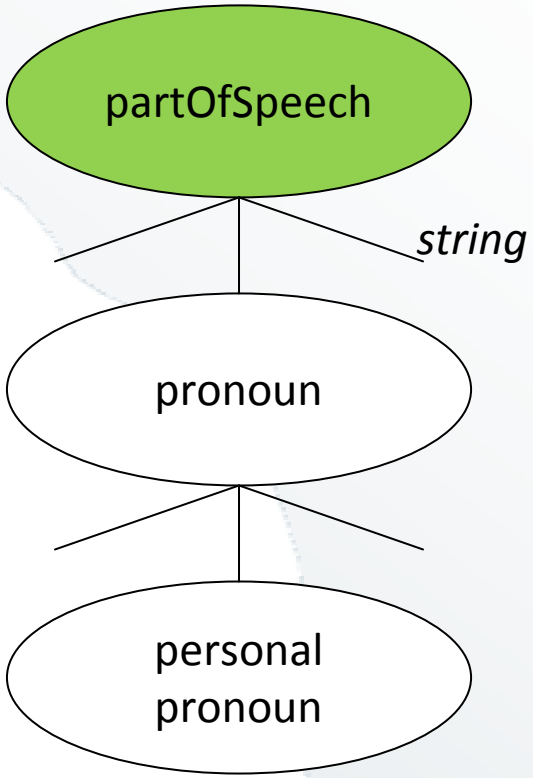
Data Category types



www.isocat.org

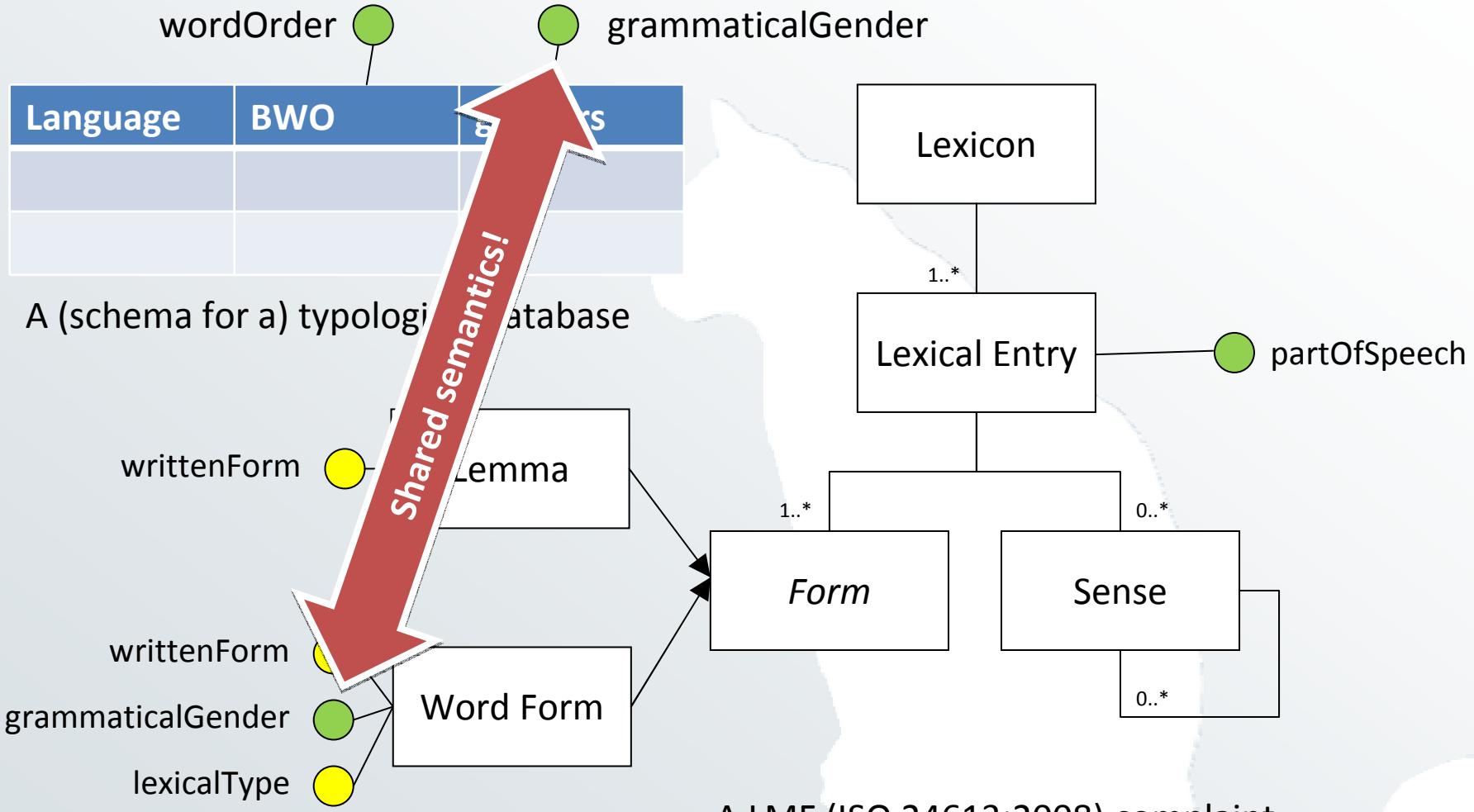
Data Category relationships

- Value domain membership
- Subsumption relationships between simple data categories
- Relationships between complex data categories are not stored in the DCR



www.isocat.org

How can you use Data Categories?



www.isocat.org

How?

```
<lmf:lexicon xml:lang="jp" alphabet="ipa">
   <lmf:entry>
        <lmf:lemma>
             <lmf:writtenForm>nihongo</…>
             …
        </…>
        …
   </…>
   …
</…>
```

ISOcat
www.isocat.org

Referencing Data Categories

- 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
- ISO/DIS 24619 Language resource management -- Persistent identification and access in language technology applications

ISOcat
www.isocat.org

Data Categories Persistent IDentifiers

- persistent identifier (PID)
  - "unique Uniform Resource Identifier (URI) that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership" (see [1])
- For Data Categories this digital object is a specific version of a Data Category specification, i.e., each version of a Data Category has its own PID

www.isocat.org

Where do you put these references?

- Preferably in a schema:

```
<rng:attribute name="alphabet"
    dcr:datcat="http://www.isocat.org/datcat/…">
  <rng:value dcr:datcat="http://www.isocat.org/datcat/…">
    ipa
  </…>
  …
</…>
```

www.isocat.org

ISO TC 37 standards using Data Categories

- Terminological Markup Framework (TMF; ISO 16642)
- Lexical Markup Framework (LMF; ISO 24613)
- TermBase eXchange (TBX; ISO 30042)
- Morpho-syntactic Annotation Framework (MAF; ISO 24611)
- Linguistic Annotation Framework (LAF; ISO 24612)

- Meta models which can be instantiated into a specific model with data categories
- However, some still refer to ISO 12620:1999 Data Categories and some don't support all types (see [3])

ISOcat
www.isocat.org

Other uses of Data Categories

- CLARIN Component Metadata Infrastructure (CMDI)
- ISO 12620:2009 provides a small XML vocabulary, DC Reference (see [4]), which provides elements and attributes to embed Data Category references in arbitrary XML documents
  - Including: XML Schema, Relax NG, TEI/ISO feature structures, …
- The references can be used in URI based ‘mappings’:
  - Including: ODD, RDF-based vocabularies (OWL, SKOS), …

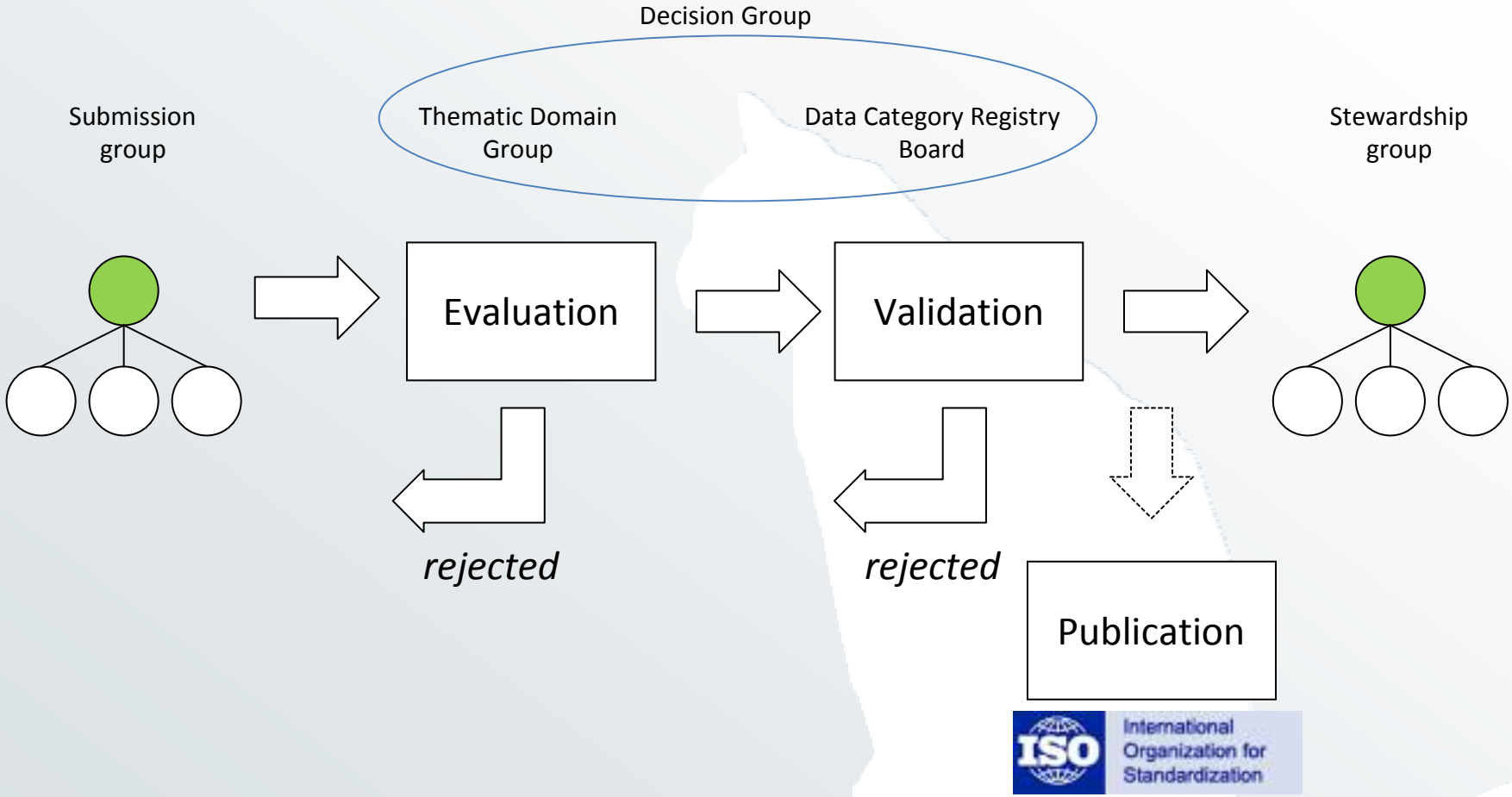
ISOcat
www.isocat.org

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

www.isocat.org

Standardize Data Categories



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

ISOcat
www.isocat.org

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 needed for your resources and share those
  - Submit (your) Data Categories for standardization
  - Free of charge

www.isocat.org

ISOcat

- Reference implementation of ISO 12620:2009
- The TC 37 Data Category Registry

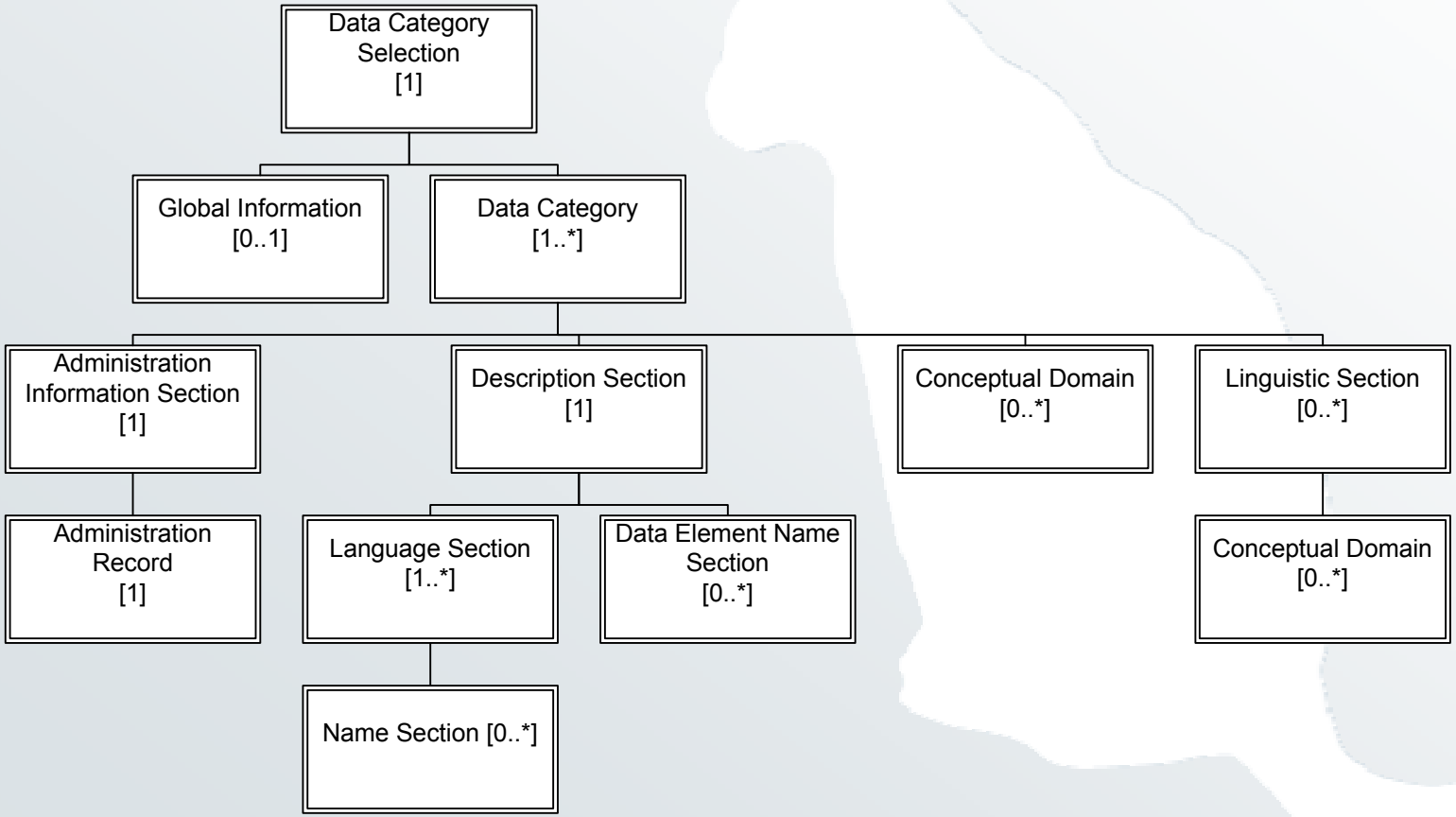
www.isocat.org

A glimpse of ISOcat

www.isocat.org

Data Category Interchange Format (DCIF)

- Simplified XML serialization of the data model (see [4])



ISOcat
www.isocat.org

RESTful Web Services

- read-only programming interface to the DCR (see [5])
- allows tools to interact with ISOcat to help an user to embed PIDs in their resources
- mainly based on DCIF
- uses authentication to access private/shared Data Categories
- currently used by:
  - LEXUS: populate an LMF model
  - ELAN: create controlled vocabularies
  - CMDI Component Editor: create concept links for component elements

ISOcat
www.isocat.org

Persistent IDentifiers

- ISOcat uses ‘cool URIs’ as PIDs (see [6])
  - these URIs will never change, but resolve to the current location in the current implementation, e.g., in ISOcat they resolve to a RESTful Web Service call
  - the isocat.org domain is bound to ISO 12620:2009 and the Registration Authority, currently the MPI, is obliged to keep the PIDs associated with this domain resolvable

ISOcat
www.isocat.org

Future work

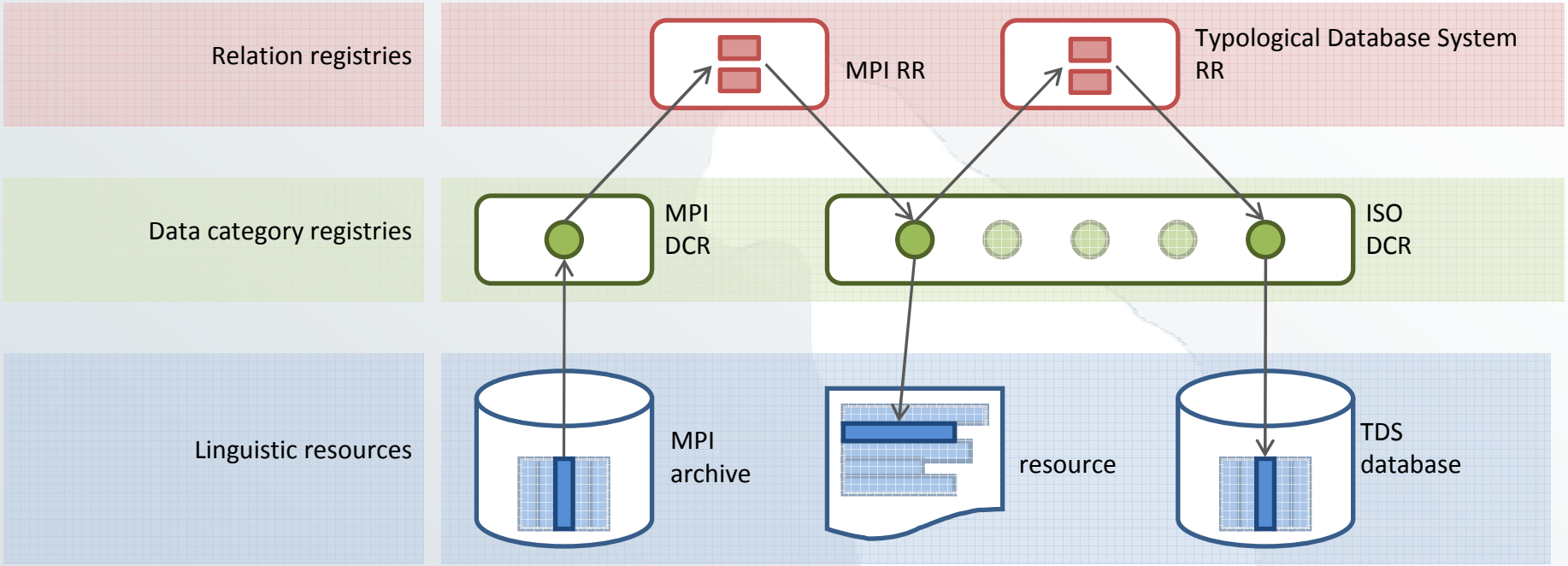
- Finish first complete version of ISOcat:
  - Standardization process
- Cleanup of the current set of Data Categories
  - TDGs cleanup their profiles
  - Standardize first sets of Data Categories
- Interaction with other TC 37 standards:
  - Migration from ISO 12620:1999
  - Full support for all types of Data Categories

www.isocat.org

More future work

- Additional Data Categories types
  - Container Data Categories
    - Complex and Simple only cover ‘leafs’ and their values
  - Data Category Concepts
    - Basic building blocks for knowledge bases
- Relation Registries
  - Stores (your) (semantic) relationships between Data Categories

Registry network



Thank you for your attention!

Visit

www.isocat.org

Questions?

www.isocat.org/forum/

isocat@mpi.nl

ISOcat
www.isocat.org

References

[1] ISO 12620, *Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources*.

[2] http://www.isocat.org/manual/DCRGuidelines.pdf

[3] M.A. Windhouwer, S.E. Wright, M. Kemps-Snijders. Referencing ISOcat data categories. In proceedings of the LREC 2010 *LRT standards* workshop. Malta, May 18, 2010.

[4] http://www.isocat.org/12620/

[5] http://www.isocat.org/rest/help.html

[6] Tim Berners-Lee, Cool URIs don't change, 1998.