ISOcat

A short introduction

Marc Kemps-Snijders\textsuperscript{a}, Sue Ellen Wright\textsuperscript{b}, Menzo Windhouwer\textsuperscript{a}

\textsuperscript{a}Max Planck Institute for Psycholinguistics, \textsuperscript{b}Kent State University

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

February 19, 2010

CLARIN-NL Bijeenkomst
ISOcat: a data category registry

• ISO 12620:2009
  – Terminology and other content and language resources — Specification of data categories and management of a Data Category Registry for language resources
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 in various working languages
  – Linguistic part
    • Conceptual domain(s for various object languages)
ISOcat 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.

At regular intervals, snapshots of the standardized subset of the DCR will be submitted to ISO.

---

**Data Category Registry (DCR)** — Core resource (All TC 37 DCS)

- Terminology
- Language coding
- Metadata for language resources
- Morphosyntax
- Private DCSs
- Individual experts
- ... etc. for other domains

**TDG Board**

- TDG terminology
- TDG morphosyntax
- TDG metadata
- TDG ...
Standardization

Submission group

Thematic Domain Group

Data Category Registry Board

Decision Group

Evaluation

Validation

Rejected

Publication

Stewardship group
Data categories and linguistic resources

A (schema for a) typological database

A (schema for a) lexicon

<table>
<thead>
<tr>
<th>Language</th>
<th>BWO</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lexicon

Lexical Entry

Form

Sense

Lemma

Word Form

writtenForm

grammaticalGender

lexicalType

partOfSpeech

wordOrder

grammaticalGender

A (schema for a) typological database

Shared semantics!
Referencing data categories

```
<dcif:dataCategory pid="http://www.isocat.org/datcat/DC-1345" type="complex">
  <dcif:administrationInformation>
    <dcif:administrationRecord>
      <dcif:identifier>partOfSpeech</dcif:identifier>
      <dcif:version>0.0.0</dcif:version>
      <dcif:registrationStatus>candidate</dcif:registrationStatus>
      <dcif:origin>?</dcif:origin>
      <dcif:creation>
        <dcif:creationDate>2004-07-09</dcif:creationDate>
        <dcif:changeDescription xml:lang="en">
          ...  
          ...
        </dcif:changeDescription>
      </dcif:creation>
    </dcif:administrationRecord>
  </dcif:administrationInformation>
  <dcif:descriptionSection>
    <dcif:profile>MorphoSyntax</dcif:profile>
    <dcif:languageSection>
      <dcif:language>en</dcif:language>
      <dcif:definitionSection>
        <dcif:definition xml:lang="en">
          Term used to describe how a particular word is used in a sentence.
        </dcif:definition>
      </dcif:definitionSection>
    </dcif:languageSection>
  </dcif:descriptionSection>
</dcif:dataCategory>
```
Annotating linguistic resources

• Schema language support for equivalence:
  – for example ODD from TEI
    <elementSpec ident="pos">
      <equiv name="partOfSpeech" uri="http://isocat.org/datcat/ISO-DC-369"/>
    </elementSpec>

• Annotation using dcr:datcat attribute:
  – for schemas or instances
  – for example RelaxNG schema
      <rng:choice>
      </rng:choice>
    </rng:element>

• XML oriented, is more needed?
Data categories as RDF resources

A domain modeling approach:

:headword a rdfs:Class .

:partOfSpeech a rdf:Property ;
    rdfs:domain :headword .

Alternative approach:

:headword a rdfs:Class .

:partOfSpeech a rdf:Class .

:hasPartOfSpeech a rdf:Property ;
    rdfs:domain :headword
    rdfs:range:partOfSpeech .

:noun a partOfSpeech .
ISOcat status

• ISOcat is under active development:
  – Now:
    • You can access public data categories and selections
    • You can create your own data categories and selections
    • You can share your data categories and selections with others (everyone, or a specified group)
  – In progress:
    • Cleanup of profiles by TDGs
    • Standardization workflow
    • Some social features (forum to discuss specific data categories)
    • Import external ‘data category’ sets, such as:
      – parts of the ISO Concept Database
      – Dublin Core
      – TEI
  – Future:
    • High availability (mirrors)
    • Relation registry
ISOcat workshop

- Utrecht, Thursday March 25, 2010
- Especially aimed at supporting Call 1 projects
- Signup at: www.clarin.nl

Program:
- A deeper introduction to ISOcat
- A tutorial on using ISOcat
- How to annotate specific linguistic resources?

Invitation

Send examples of the types of linguistic resources your project wants to annotate with data category references to

isocat@mpi.nl

and we will discuss them at the workshop!
Thank you for your attention!

Visit
www.isocat.org

Questions?
isocat@mpi.nl