ISOcat and the Typological Database System

Menzo Windhouwer
(MPI for Psycholinguistics)
Standards: What purpose do they have?

- What goal does the MPI archive have?
  - “... the aims of the DOBES programme are to document languages in their cultural setting and present them in such a way that
    - they are useful for linguistics, anthropology, history, comparative literature, and other disciplines; ...”
  - “All material has to be stored in digital form according to open standards, and the material should be accessible from all over the world via the Internet. In addition, the archive has to take care of the long-term persistency of the digital material, which is not evident given the limited life-time of our state-of-the-art storage media.”

(http://www.mpi.nl/DOBES/dobesprogramme/)
Standards help us to realize these goals

- Syntax (generic)
  - Unicode
  - XML
  - ...
- Structure (domain specific)
  - LMF
  - LAF/MAF/GrAF
  - EAF
  - ...
- Semantics (sometimes even linguist specific 😊)
  - DCR
  - GOLD
  - ...

The more specific we get the fewer standards there are, so to meet our goals we try to help to create (flexible) standards in these areas.
Official standardization bodies

• International:
  – ISO
  – OASIS
  – W3C (recommendations)
  – …
• Regional (Europe):
  – CEN
  – …
• National:
  – NEN
  – DIN
  – …
• But actually anybody can (try to) create a *de facto* standard
Standards we are/were involved in

- Lexical Markup Framework (LMF; ISO 24613:2008)
  - LEXUS is a LMF reference implementation
- Data Category Registry (DCR; ISO 12620: 2009)
  - ISOcat is the DCR reference implementation
  - MPI is the Registration Authority for the TC 37 DCR
- Language resource management -- Persistent identification and access in language technology applications (CitER; ISO/DIS 24619:2009)
- ELAN Annotation Files
- [http://www.clarin.eu/recommendations](http://www.clarin.eu/recommendations)
- ...

DGfS-CNRS Summer School on Linguistic Typology

8/20/2010
What will follow?

• ISO 12620:2009
  – the Data Category Registry ISOcat
  – MPI is the registration authority
    • develops and hosts ISOcat
  – an increase number of tools interact with ISOcat

• Typological Database System
  – uniform interface to a dozen typological databases
  – shows how typological resources could benefit from ISOcat (to be realized in the TDS Curator project)