Towards an ISO standard representation for proper names

Béatrice Bouchou
beatrice.bouchou@univ-tours.fr
LI – Université François Rabelais de Tours
16th April 2007
Summary of the presentation

- ISO/TC 37
  - Data categories
  - Terminological Markup Framework (TMF)
  - Lexical Markup Framework (LMF)

- Prolexbase
  - in TMF
  - in LMF
  - XML (?)
What are ISO standards for?

- Provide a common model for the creation and use of NL resources
- Manage the exchange of data between and among these resources
- Enable the merging of electronic resources

Range
- Monolingual
- Bilingual
- Multilingual
- Linguistic description range from morphology, syntax, semantics to multilingual representation
- Languages are not restricted to European languages
- The range of targeted NLP applications is not restricted

Scalability
- Same specifications for both small and large lexicons
ISO/TC 37

- Terminology and other language resources
  - **SC3** - Computer applications in terminology
    - ISO 12200 - Martif
      - Latest version of TEI Terminology chapter
    - ISO 12620 - Data categories (under revision)
    - ISO 16642 - TMF (Terminological Markup Framework)
  - **SC4** - Language Resource Management
ISO/TC 37/SC4

The following standards are under preparation:

- ISO/NWI 21829  Terminology for language resources
- ISO/NP 23679-1  Word segmentation of written texts – Part 1: General principles and methods
- ISO/CD 24610-3  Language resource management – Feature structures – Part 3: Word segmentation for other languages
- ISO/WD 24611  Language resource management – Morphosyntactic annotation framework
- ISO/WD 24613  Language resource management – LMF (Lexical Markup Framework)
Two-level standards:

- the **high level** specifications provide structural elements, i.e. classes and relations between them: the meta-model

- the **low level** specifications provide standardized constants, i.e. data categories used to “adorn” the classes: ISO 12620
Data categories

- **Definition**
  - Feature names and values used to describe natural language resources

- **Example**
  - Features: /part of speech/, /grammatical gender/
  - Values: /feminine/, /plural/, /dual/, /ablative case/

- **Role**
  - Characterization of structural elements (specification)
  - Linguistic range identification (documentation)
**Example of DCR**

**DCRegistry**

**version**: 1

<table>
<thead>
<tr>
<th>Administration Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC ID</td>
</tr>
<tr>
<td>DC Name</td>
</tr>
<tr>
<td>DC Definition</td>
</tr>
<tr>
<td>DC Source Comment</td>
</tr>
<tr>
<td>Data Type</td>
</tr>
<tr>
<td>Level(s)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Administration Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC ID</td>
</tr>
<tr>
<td>DC Name</td>
</tr>
<tr>
<td>DC Definition</td>
</tr>
<tr>
<td>Concept-related Comment</td>
</tr>
<tr>
<td>Data Type</td>
</tr>
<tr>
<td>Level(s)</td>
</tr>
</tbody>
</table>
## Example of DCR

### Conceptual Domain

<table>
<thead>
<tr>
<th>DC ID</th>
<th>Conceptual Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Name</td>
<td>Conceptual Domain</td>
</tr>
<tr>
<td>DC Definition</td>
<td>[from-en] On the DS level, this field is used to relate the category under description with the set of all its possible values (expressed as a list of data categories). When necessary a datatype (in the sense of XML schemas) may be provided instead of a list of values [from-en] On the LS level, to be used when a data category is to be associated to a specific subset of the values declared at DS level</td>
</tr>
<tr>
<td>DC Source Comment</td>
<td>ISO 11179-3</td>
</tr>
<tr>
<td>Data Type</td>
<td>basicText (open)</td>
</tr>
<tr>
<td>Level(s)</td>
<td>Language Section, Description Section</td>
</tr>
</tbody>
</table>

### Creation Date

<table>
<thead>
<tr>
<th>DC ID</th>
<th>Creation Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Name</td>
<td>Creation Date</td>
</tr>
<tr>
<td>DC Definition</td>
<td>[from-en] the date when the data category has been initially created (for instance in an expert's working space/private area);</td>
</tr>
<tr>
<td>Concept-related Comment</td>
<td>[psb] must be refine by /Change description/ to show the modifications between the last version and the current one.</td>
</tr>
<tr>
<td>Data Type</td>
<td>basicText (open)</td>
</tr>
<tr>
<td>Level(s)</td>
<td>Administration Record</td>
</tr>
</tbody>
</table>
### Example of DCR

<table>
<thead>
<tr>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DC ID</strong></td>
</tr>
<tr>
<td><strong>DC Name</strong></td>
</tr>
<tr>
<td><strong>DC Definition</strong></td>
</tr>
<tr>
<td><strong>DC Source Comment</strong></td>
</tr>
<tr>
<td><strong>Data Type</strong></td>
</tr>
</tbody>
</table>
| **Level(s)** | Language Section  
Description Section  
Name Section  
Used only for refinement |
Data categories

- Data Category Registry (DCR): http://syntax.inist.fr
  - NLP resource developers can refer to this DCR, while building their proper DCR…

- Motivation
  - Reference framework for comparing models and structures for NLP
  - Towards a better interoperability between information systems
Example: Morphalou

- Dictionary of inflected forms for French (www.cnrtl.fr)
- Only forms yet (no senses)
  - 60,940 common nouns
  - 8,790 verbs
  - 22,790 adjectives
  - 1,579 adverbs
  - etc.
Data Categories in Morphalou

- <orthography>
- <grammaticalCategory>
- <grammaticalGender>
- <grammaticalNumber>
- <spellingVariantOf>
- <feminineVariantOf>
- <originatingEntry target="68340">
- …
Meta-models

- TMF
  - standards and guidelines for creating and using terminological data collections
  - Core model + data categories

- LMF
  - creation and use of electronic lexical resources
  - generic skeleton + data categories
  - More components in the skeleton…
TMF meta-model

- All following slides on TMF are taken from:

  Laurent Romary  
  Laboratoire Loria-INRIA
TMF Meta-model

Terminological Data Collection (TDC)

*  

Global Information (GI)

*  

(Concept) Terminological Entry (TE)

*  

Complementary Information (CI)

Language Section (LS)

*  

Term Section (TS)

*  

Term Component Section (TCS)

Figure: L. Romary, G. Budin

Prolex
Example in TBX [www.lisa.org]

```xml
<termEntry id="ID67">
  <descrip type="subjectField">manufacturing</descrip>
  <descrip type="definition">A value between 0 and 1 used in ...</descrip>
  <langSet lang="en">
    <tig>
      <term>alpha smoothing factor</term>
      <termNote type="termType">fullForm</termNote>
    </tig>
  </langSet>
  <langSet lang="hu">
    <tig>
      <term>Alfa ...</term>
    </tig>
  </langSet>
</termEntry>
```
TMF model of previous TBX example

```
TE
  id='ID67'
  subjectField='manufacturing'
  definition='A value...'

LS
  lang='en'

TS
  term='alpha smoothing factor'
  termType='fullForm'

LS
  lang='hu'

TS
  term='...'
```
Concept view: example

- subject: appearance of materials
- definition: degree of obstruction to the transmission of visible light
- terms
  - term 1: opacity
    - POS: n
  - term 2
    - ...
  - encyclopedic notes
TMF multilingual

Language 1

Language 2

Language 3

Concept

Definition

Term 1  Term 2  Term...

Usage  Context  ...

Term...
LMF skeleton

- All following slides on LMF are taken from:

  Monica Monachini
  CNR-ILC - Pisa
Structure of LMF

Structural skeleton, with the basic hierarchy of information in a lexical entry

extend a subset of core-model classes; are conformant to the core model; cannot be used regardless to the core model.

Core Package

- NLP Morphology extension
- NLP Inflectional paradigm extension
- NLP MWE pattern extension
- NLP Syntax extension
- NLP Multilingual notations extension
- NLP Semantic extension
- MRD extension
Core package

Prolex

Form consists of a text string that represents a single word or a multi-word expression.

One to many Representation Frames can be associated with Form, each of which contains a form and data categories that specify the orthographic types and name of the word.

Container for managing the top level language components. Homonyms are distinct LE if their morphologies are distincts.

Cross-reference pivot that can link to many Lexical Entries within or across Lexicons.

Sense specifies or disambiguates the meaning and context of a form.
Package for extensional morphology

1st strategy: describe the morphology representing explicitly all inflections
Package for inflectional paradigm

2nd strategy:
declare an inflectional paradigm;
use the inflectional paradigm extension for defining it
Package for NLP syntax

**Syntactic behavior** represents one of the behaviors of one (or more) senses

**Construction** describes one syntactic construction and can be shared by all words with the same syntactic behavior

**Self** refers to the head lexical entry and describes syntactic properties

**Syntactic Argument** describes a syntactic actant

**ConstructionSet** regroups together various **Syntactic Constructions** and factorizes syntactic descriptions to have a minimum of syntactic behavior elements in the lexicon.

---

**Description of Syntactic Argument**

- Function: subject
- Syntactic Constituent: NP

**Description of Construction**

- ID: amare-SyntFrame
- Self ID: amare-self
- Auxiliary: avere
Package for NLP semantics

Predicative Representation describes the link between Sense and Semantic Predicate

Semantic Predicate describes an abstract meaning

Semantic Argument describes a semantic actant and is linked with its syntactic counterpart
Example (package for semantics)
Package for Multilingual representation

Sense 0..* Sense Axis 0..* 0..* 1
SynSet 0..* Sense Axis Relation 0..* 1 0..1
Sense Axis Relation 0..* Target Test 0..*
Source Test 1 1 0..*
Syntactic Behavior 0..* Transfer Axis 0..*
Transfer Axis Relation 1 0..* 1
SenseExample 0..* 0..* Example Axis 0..* 0..*
Package for Multiword expressions

**List of Components**
- 0..1
- 0..*
- 1

**Lemmatised Form**
- 1
- 0..*
- 0..1

**MWE Pattern**
- 1
- 0..*

**Combiner**
- 1
- 0..*

**Combiner Argument**
- 0..*
- 1
Example (multiword expression)

: List of Components

: Lemmatised Form
  writtenForm = throw

: Lemmatised Form
  writtenForm = to

: Lemmatised Form
  writtenForm = the

: Lemmatised Form
  writtenForm = lion

: MWE Pattern
  id = VP Somebody PP
  comment = for a pattern, VP somebody IndirectObject

: Combiner
  head = true
  constituent = VP
  rank = 0
  graphicalSeparator = space

: Combiner Argument
  function = directObject

: Combiner Argument
  function = indirectObject

: Combiner
  constituent = NP
  semanticRestriction = human

: Combiner Argument
  rank = 1
  graphicalSeparator = space

: Combiner Argument
  rank = 2
  graphicalSeparator = space

: Combiner Argument
  rank = 3
  graphicalSeparator = space

: Combiner
  constituent = PP
  number = plural
TMF and LMF (L. Romary)

- Terminology (and ontology)
  - Term: corresponds to a concept in a specific domain
  - Concept → several possible linguistic forms
    - Onomasiological view

- Lexicography
  - Lexical entry + genericity
  - LE → one or several senses
    - Semasiological view
In general, a terminological resource is organized into **concept entries**, each of which includes one or more terms designating a particular concept.

In general, a lexical resource is organized into **lexical entries**, each of which includes one or more senses of a particular lexical item (a word or phrase).

A concept entry containing multiple terms can be split into multiple lexical entries, one per term...

Multiple lexical entries associated with the same concept can be combined into one concept entry...
Prolexbase

- A database of proper names and their relationships
- Four levels in the conceptual model:
  - Instances
  - Linguistic
  - Conceptual
  - Meta conceptual
Details on Prolex model

See the presentation of Denis Maurel in the same seminar...
Prolex features

- Proper name “sense”: a point of view on a referent
- Semantic description is at multilingual level
- The multilingual level is important (it is false that “a proper name do not need to be translated”)
- The prolexeme groups several terms, these terms are related to the same pivot (“sense”)
Prolex features

- Importance of *derivations performed from the prolexeme*:
  - derivatives are related to the same point of view on the same referent as their prolexeme

- The *complete linguistic description* (including morphology, syntax, and semantics):
  - extensional (the instance layer)
  - intensional (inflection paradigms, rules for aliases and derivatives).
ISO Standard for Prolex

• LMF?
  • Proper names are not a « specialized language » (as a terminology)
  • Complete linguistic description
  • But special « sense », no semantic description at lexical entry level
  • But derivative hierarchy not easy to represent in LMF
ISO Standard Prolex: Meta-model

- **TMF ?**
  - Proper names are special citizens of languages: *pivots resemble concepts*
  - Semantics defined at concept level
  - Multilingual links via concepts
  - **But**: complete lexical information not easy to express
  - **But**: derivative hierarchy difficult to represent in TMF
Prolex in LMF (linguistic – instance)

- **Prolexeme → Lexical Entry**
  - the *Sense* part of this LE is « prolexeme » (+ number)
- **Alias → Form** of the LE that represents the corresponding prolexeme
  - with the Related Form class of Morphology extension
- **Derivative → Lexical Entry**
  - the Sense part of this LE is the relation name (e.g. « relational adjective ») + its « parent » number
- **Links of « derivation » (from « parent » to « child »): Entry Relation (?)**
- **Instance → Form** of the LE that represents the corresponding prolexeme or derivative
  - with the subclasses Lemmatised Form and Inflected Form
Prolex in LMF (conceptual – meta)

- **Pivot** → **Sense Axis** of Multilingual extension
  - Links prolexemes (via their sense part)

- **Conceptual relations** → **Sense Axis Relation** of Multilingual extension + data categories

- **Type** → **Sense Axis Relation** + Interlingual External Ref of Multilingual extension.

- **Language name** → **Lexicon Information** part

- **Multilingual links** → **Axis** of Multilingual extension
Prolex LMF Abstract Model

- Lexicon (fr)
- Lexicon (en)
- LE (prolexeme)
- LE (derivative)
- Forms
- Sense
- Syntactic behavior
- Sense Axis
- Typology
- Synonymy
- Database

Syntactic behavior

Sense

Forms
Prolex in **TMF** (linguistic – instance)

- Each instance of prolexeme → Term Section (TS)
- Each instance of derivative → Term Section (TS) under the same TE as their prolexeme

A data category (\eg "associative relation") can link the prolexeme with its derivatives (and derivatives with their derivatives)

Data categories: "grammatical gender", "grammatical number" are associated to TS

- **Alias** → Data category attached to the TS of the prolexeme: "abbreviated form", "acronym", "variant", etc.
Prolex in TMF (conceptual – meta)

- **Pivot → Terminological Entry (TE)**
- **Conceptual relations → Data categories:** "related concept", "concept relation" …
- **Type → Data categories:** "subject field", "broader concept"…
- **Language → Language section (LS)**
- **Multilingual links → Via the TEs, plus data categories attached to the LS instances**
Prolex TMF abstract model

Related concept synonym

TE (pivot)

Terminology

Subject field

LS (en)

Related concept synonym

LS (fr)

TS (prolexeme)

Abbreviate form

Variant

gender

number

relation

TS (derivative)

Subject field

LS (en)

TS

number

relation

Subject field
Prolex TMF example

- **Existence**
  - historical

- **Terminology**
  - Pivot 46929

- **Subject field**
  - country

- **LS (en)**
  - United States of America

- **Prolexeme**
  - United States of America

- **Abbreviate form**
  - United States

- **Acronym**
  - USA

- **Number**
  - plural

- **Relation**
  - relational adjective

- **LS (fr)**
  - Etats-Unis d'Amérique

- **Derivative**
  - American
What about XML?

- XML design is not so simple
- XML design is not mature
- Proposal in NLDB 2005
  - Database is (i) a multilingual part and (ii) a set of languages (not so far from LMF…)
  - Main entries in multilingual part: pivots
  - Main entries in language part: prolexemes
- From conceptual (abstract) model to a logical model: many solutions for XML…
Representing relations*:

- Polysemeia
  - LMF: several senses of a lexical entry
  - TMF: not considered (a term is related to only one concept…)
  - Prolex: as TMF, one pivot for each point of view on a referent

- Synonymy
  - LMF: Links from one LE (one sense) to another one
  - TMF: in concept structure*
  - Prolex: conceptual relations between pivots

(*L. Romary)
XML Instances of ISO models

- I did not find any report on discussions

- XML « instances » of TMF: TML
  - MSC (MARTIF with Specified Constraints)
  - Geneter
  - GMT: a tool for comparing two TML...

- XML « instances » of LMF:
  - XML schemas: a DTD + a RelaxNG specification
  - To be analysed
Conclusion

- A lot of work remains to be done…