Terminological Data Modeling
Traditional and New Aspects

TKE 2010 – Workshop on ISO/CDB
14th August 2010 – Dublin

Klaus-Dirk Schmitz
Fakultät 03 - ITMK
Fachhochschule Köln
klaus.schmitz@fh-koeln.de
Content

- General principles of terminology data modeling
  - Data categories
  - Meta model
  - Modeling principles
- Practical implementations and improvements
  - Concept orientation
  - Term autonomy
  - Locale problem
  - (Non-linguistic representations)
- New aspects and outlook
- Conclusion
Data Modeling

- A data model is an abstract model that describes how data are represented and accessed. Data models formally define **data elements** and **relationships among data elements** for a domain of interest.

- Data elements = data categories: ISO 12620 (ISOcat)

- Relationships = meta model: ISO 12200 / ISO 16642 (TMF)

- A **terminological meta model** is an abstraction of data models for terminological applications (data bases, formats etc.) highlighting the common properties of these applications.
Terminological meta model

Terminological Data Collection

- Global Information Section
- Concept/Entry
  - Language Section
    - Term Section
      - Term Component Section
- Complementary Info
Argumentation for TMF

How does the common Meta Model look like?

LU = Linguistic Unit (Term, Phrase, Segment)
But we need multilinguality!
Argumentation for TMF

- How does the common Meta Model look like?

- But we need synonyms!
Argumentation for TMF

Metadata

Lang. 1

LU 1
German

LU 2
German

LU 3
English

Lang. 2

LU 1
English

LU 2
English

Lang. 3

LU 1
Spanish

LU 2
n.n.

Lang. n

LU 1
n.n.

LU 2
n.n.
Modeling Principles: Concept Orientation

- Terminological data modeling has to follow the principle of **concept orientation**: a terminological entry has to contain all terminological data related to one concept (all languages, all terms, all related information).

- The meta model (and ISO 704 + 1087-1) demands this.

- The terminological entry is a logical unit, not a physical one (≠ record).
Modeling Principles: Term Autonomy

- Terminological data modeling has to follow the principle of **term autonomy**: all terms representing the concept (including synonyms, abbreviated forms and orthographic variants) can be retrieved and documented with all necessary term-related data categories.

- Terminological meta model: Unlimited number of term sections, but only one term in each section.
Modeling Principles

Concept represented by ID-No. and/or classification / notation

Language 1
Term 1 + AuxInfo
Term 2 + AuxInfo

Language 2
Term 1 + AuxInfo
Term 2 + AuxInfo
Term 3 + AuxInfo

Language 3
Term 1 + AuxInfo
Terminological Meta Model (12200)
Entry Structure

Create an entry structure for your termbase entries by specifying the level at which descriptive fields are used. Specify field settings if required.

Entry structure:
- Entry level
- Index level
- Term level

Available descriptive fields:
- Context
- Definition
- Grammatical Gender
- Grammatical Number
- Illustration
- Note
- Part of Speech
- Source
- Status
- Subject Field
- Term Type
- Usage Register

Field settings:
- Mandatory
- Multiple

© K.-D. Schmitz
Bei der elektrischen Nutzbremse werden die Fahrmotoren als Generatoren betrieben und die dabei gewonnene Energie in das Netz zurückgeleitet.

Source: http://www.hochgeschwindigkeitszuege.com, 14.11.2006

**Regenerative Brake**

A braking system which enables an electric locomotive or train to reduce its energy consumption by feeding back the traction supply power generated by the motion of the train when it is descending a gradient.

Source: http://www.mda.org.uk, 8.10.2006
Language vs. locale:

- In computing and software localization, a **locale** is a set of parameters that defines the user’s country, language, keyboard layout, writing system, date format etc.

- A locale identifier consists of at least a language ID and a region/country ID.
Language vs. locale:

- Locale can be used in terminology management to describe that different terms within the “same” language are used in different geographical regions to express the same concept.

**Example:**

- English: windscreen <UK>
- English: windshield <US>
Practical Implications of Terminological Data Modeling

Language vs. locale:

- **Traditional way:**
  - use the language level as in the meta model
  - manage regional variants of the term as synonyms
  - use data category for geographical usage (ISO country code) as attributes to the term
winds

English

German

Default input model

Entry level
Entry number: 295
Project Code: Sample entry for TSTT06
Domain: automotive (TRA)

EN

English
windscreen
Part of Speech: Noun
Regional Label: UK

windshield
Part of Speech: Noun
Regional Label: US

German
Windschutzscheibe
Gender: f
Part of Speech: Noun
Language vs. locale:

- Other approach (e.g. as in MultiTerm):
  - index level (=language level) can be specified by a combination of language and country (=locale)
  - terms of a specific locale are managed in separate language/locale section
  - searching for terms is difficult if many terms are the same in all regions (as for US and UK English)
  - but allows for separate definitions / figures for each locale
Index Fields
Select the languages you wish to include in your termbase. Accept the default index field label for each language or customise it to suit your requirements.

Languages:
- Afrikaans

Available index fields:
- DE  German (Germany)
- DE-AT  German (Austria)
- DE-CH  German (Switzerland)
- EN-GB  English (United Kingdom)
- EN-US  English (United States)
- FR  French (France)
- FR-BE  French (Belgium)
- FR-CA  French (Canada)
- FR-CH  French (Switzerland)

Sort order:
- Case-sensitive
- Ignore non-alphabetic characters

Field label:
Entry level
Entry number: 4
Project Code: Sample entry for TSTT06

**English (United Kingdom)**
Non-ling. Repres.

**plug**
PartOfSpeech: noun

**English (United States)**
Non-ling. Repres.

**plug**
PartOfSpeech: noun

**French (Switzerland)**
Non-ling. Repres.

windscreen windscreen plug
French (Switzerland)
Non-ling. Repres.

fiche
Gender f

German (Germany)
Non-ling. Repres.

Stecker
Gender m
PartOfSpeech noun

Stecker

German (Switzerland)
Non-ling. Repres.
Practical Implications of Terminological Data Modeling

Language vs. locale:

- Such a locale (language+country) solution will also support data modeling for distributed / federated terminological applications.

- Instead of a region, an organization or a company can be seen as a locale, e.g. ISO, ANSI, BSI as locales

- Application: ISO/CDB for ISO and its member bodies

- Do we need a Locale Section instead of the Language Section in our meta model? What are the consequences?
Sample entry: not so many term-related datCats! term autonomy? what is locale-specific?

**Term:** part of speech

**Definition:** A category assigned to a word based on its grammatical and semantic properties. PERMISSIBLE INSTANCES: Examples of parts of speech commonly documented in terminology databases can include: a) noun b) verb c) adjective
Term: grammatical gender

Definition: A grammatical category that indicates grammatical relationships between words in sentences. PERMISSIBLE INSTANCES: Types of grammatical gender commonly documented in terminology databases include: a) masculine b) feminine c) neuter d) other

Example(s):
1. In French, vie (life) is feminine and is used with feminine articles such as la, the feminine pronoun elle, and feminine adjective endings, e.g., une vie longue.

Note: The concept of gender varies from language to language and is not a universal feature of all languages.
ISOcat data model
Conclusion

- ISO standards provide a set of profound principles, methods, and formalisms for terminological data modeling.

- Improvements in the fields shown will make ISO results more suitable for advanced terminological applications.

- The proposed adaptations / modifications will enlarge the scope of applicability to data modeling for other kinds of structured content, particularly for multilingual and multicultural applications.
Thank you for your attention
Questions?

Prof. Dr. Klaus-Dirk Schmitz
Fachhochschule Köln
Fakultät 03 - ITMK/IIM
Mainzer Str. 5
50678 Köln
klaus.schmitz@fh-koeln.de