Terminology Management – Data Categories & Modeling Principles

Terminology Summer School - Vienna
09 - 13 July 2012

Klaus-Dirk Schmitz
Institute for Information Management
Faculty 03
University of Applied Sciences Cologne
klaus.schmitz@fh-koeln.de
Designing a terminology management solution

Before designing a terminology management solution and choosing, adapting or programming a terminology management application:

- Analyze the needs and objectives
- Specify the user groups, tasks and workflow
- Define the terminological data categories needed
- Take into account the basic modeling principles
- Model the terminological entry
- Select, adapt, develop the software

Meta data
ISO/TC 37 Definitions

- Data elements and categories are strictly defined:
  - **data element**
    - unit of data that in a certain context is considered indivisible
    - one unit of information in a data entry
  - **data category** (data element type)
    - result of the specification of a given data field

2) Data Categories
The wardrobe model

From unorganized chaos ...
The wardrobe model

... to useful structured organization
Data modelling for termbanks

The Time Manager
Manage Your Wardrobe System

Save time!
Save money!
Get immediate results!

Are you frustrated with having nothing to wear? You CAN look well put together with a minimum of effort and save money!

Learn how to:
* Spend less and have more
* Get dressed in minutes -- and look great!
* Pack lightly for a weekend trip or a long vacation

Other wardrobe books tell you how to dress on a big budget. This is a realistic guide for today's busy woman. You'll not only look pulled together but you'll spend less too!

K.-D. Schmitz, IIM, FH Köln
Data modelling for termbanks
Data categories

- First comprehensive analysis of terminological data categories used in term bases for the preparation of ISO 12620 and ISO 12200

- First standard for data categories: ISO 12620:1999

- Completely new version of ISO 12620:2009

- Data Category Registry ISOcat (www.isocat.org)
  The Registration Authority of the TC 37 DCR is the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Data categories (ISO 12620: 1999)

A.2.2.1 part of speech

NONADMITTED TERM1: grammatical category
NONADMITTED TERM2: word class

DESCRIPTION: 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
Data categories (ISOcat)

Terminology
- Applications would normally collect their own data category selections as manageable subsets of this profile. The profile includes term-related data categories such as those for points of grammar and register, us concept-related and administrative data categories. Terminology in this sense differs from controlled vocabulary and classification systems commonly used for document and object retrieval, although there are similar writing definitions. The profile is likely to include a number of Data Category Selections, for instance:
  - The TBX family of TermBase eXchange formats (including TBX-Basic) compliant with ISO 30042:2008
  - The ISO/IEC Format for terminological definitions used in ISO/IEC standards (ISO 10241)
  - Translation-oriented terminology management as reflected in ISO 12616:2002

Forum
- public forum

Status
- established

Activity
- active

Members
- Chair: Wright, Sue Ellen (Kent State University)
- Member: Aquado de Cea, Guadalupe (AENOR, Universidad Politécnica de Madrid)
- Member: Broeder, Daan (MPI for Psycholinguistics)
- Member: Budin, Gerhard (University of Vienna)
- Member: Francoppeulo, Gil (Tagmatica INRIA Spotter / Paris)
- Member: Karsch, Barbara Inge (BK Terminology)
- Member: Le Meur, Andre (Universite Rennes 2)
- Member: Lommel, Arle (LISA)
- Member: McKay, Helen (Togletext)
- Member: Melby, Alan (BYU)
- Member: Pareja-Lora, Antonio (Universidad Complutense de Madrid (UCM))
- Member: Popiölek, Monika (PKN_MAart)
- Member: Reinke, Detlef (DTT)
- Member: Schmitz, Klaus-Dirk (Cologne University of Applied Sciences)
- Member: Warburton, Kara (TC37, SC3)
- Member: Wittenburg, Peter (MPI for Psycholinguistics)
Data categories (ISOcat)

1.1 Administration Record

Identifier: partOfSpeech
Version: 1.0
Registration Status: private
Administration Status: private
Justification: Standard, frequently required data category in terminology management, lexicography, morphology, and resources.
Origin: ISO 12620
Explanatory Comment: ISO12620A-020201
Explanatory Comment: The value domain for part of speech varies with communities of practice and intentional aims of various approaches.
Unresolved Issue: We need to harmonize issues involving different DCs for different TDGs and groups in order to ensure interoperability.
2.2 Data Element Name Section

Data Element Name: part of speech
Source: ISO 12620; ISO 30042

2.3 English Language Section

Language: English (en)

2.3.1 Name Section

Name: part of speech
Name Status: preferred name

2.3.2 Name Section

Name: pos
Name Status: admitted name

2.3.3 Definition Section

Definition: A category assigned to a word based on its grammatical and semantic properties.
Source: ISO 12620

2.3.4 Example Section

Example: noun
Source: Mitre; TEI (green text); 1951

2.4 German Language Section

3. Conceptual Domain

Data Type: string
Profile: Terminology
Value: /adjective/ → adjective
Value: /adverb/ → adverb
Value: /noun/ → noun
Value: /properNoun/ → proper noun
Value: /verb/ → verb
Data Categories

- Analyze in detail which type of information you have to deal with

- Consult ISO 12620 (data categories) for getting ideas and matching your needs with standardized categories

- Not only important for an adequate and perfect solution but also for future data exchange with others
Typology of data categories I

- **Complex data categories**
  - **Open data categories**
    content not predictable and defined by specification
    e.g.: *term, definition, note*
  - **Closed data categories**
    content defined by a limited set of possible values
    e.g.: *gender, part of speech, geographical usage*

- **Simple data categories**
  content only *yes* or *no*; values of closed data categories
  e.g.: *masculine, noun, DE*
Typology of data categories II

- Concept-oriented data categories
e.g.: subject field, figure

- Language-oriented data categories
e.g.: definition

- Term-oriented data categories
e.g.: part of speech, context

- Administrative data categories
e.g.: author, date, note

- Special data categories
e.g.: term, language, (structural elements), (shared resources)
Granularity of data categories

- The degree of detail that can be achieved by using the available data categories to document terminological information

- Low granularity:
  e.g.: $grammar = m,n,s$ (masculine noun singular)

- High granularity
  e.g.: $part \ of \ speech = noun$
  $grammatical \ gender = masculine$
  $number = singular$

Advantage: better retrievability
Elementarity of data categories

- Only one of a *thing* can occupy a data element
  - e.g., only one term in a term field

- Only one *kind of thing* can occupy a data element
  - e.g., no terms or synonyms listed as such in definition fields
Elementarity of data categories

- **Error:**
  - en:term = United Nations (UN)

- **Correct:**
  - en:term = United Nations
    - term type = full form
  - en:term = UN
    - term type = acronym

- Combinability enables us to identify individual units of content: /term/ combines here with /term type/.
Elementarity of data categories

Error:
- definition = international organization that ... (Merriam Webster, 10th. Edition 2004, p. 256)

Correct:
- definition = international organization that ...
- source = *Webster2005*, p. 256
- *Webster2005* points to a shared resource (bibliographical entry)

Combinability: Source can be used with a term or any text or graphics field.

4) Elementarity
The same data category can be modeled in multiple ways (some of them not too bright!)

- **gender**
  - value = m, f, n
  - value = masculine, feminine, neuter

- **gender**
  - **masculine** = yes/no
  - **feminine** = yes/no
  - **neuter** = yes/no

6) **Variance**
Data modeling variances

Complex example

- **term**: ink jet printer
  - superordinate concept: non-impact printer
  - subordinate concept: bubble jet printer
  - coordinate concept: laser printer

- **term**: ink jet printer
  - related concept: non-impact printer
    type of relation: superordinate
  
  related concept: bubble jet printer
    type of relation: subordinate
  
  related concept: laser printer
    type of relation: coordinate
Concept orientation

- All terminological information belonging to one concept including all terms in all languages and all term-related and administrative data must be stored in one terminological entry.

concept = terminological entry
Lexicographical view / model / entry
Terminological view / model / entry

descriptive terminology management
Terminological view / model / entry

term

(concept)

(term)

(term)

(term)

prescriptive terminology management

K.-D. Schmitz, IIM, FH Köln
ribbon

Pronunciation
- Audio (US) file
- Rhymes: -iban

Etymology
From Old French riban (French: ruban)

Noun
ribbon (plural ribbons)
1. A long, narrow strip of material used for decoration of clothing or the hair or gift wrapping.
2. An inked ribbon against which type is pressed to print letters in a typewriter or printer.
3. [computing, graphical user interface] A toolbar that incorporates tabs and menus.

Translations
- long, narrow strip of material
- inked ribbon

See also
- riband

Verb
to ribbon (third-person singular simple present ribbons, present participle ribboning, simple past and past participle ribboned)
1. to decorate with ribbon

Synonyms
- beribbon
Ribbon (disambiguation)

From Wikipedia, the free encyclopedia

Ribbon may refer to;

- Ribbon (award), a term for an award.
- Ribbon (computing), user interface concept.
- Ribbon (group), a Japanese J-pop group which consist of Hiromi Nagasaku, Arimi Matsuno and Aiko Satoh.
- Ribbon bar, small devices worn by military, police, fire Service personnel or by civilians.
- Ribbon cable, a cable with many conducting wires running parallel to each other on the same flat plane.
- Ribbon, a monthly Japanese shōjo manga magazine.
- Ribbon, typewriter an inked band of fabric used for typewriters, receipt printers and dot-matrix printers
- Awareness ribbon a ribbon worn to signify sympathy for, and raise awareness of, a cause espoused by the wearer

This disambiguation page lists articles associated with the same title. If an internal link led you here, you may wish to

Categories: Disambiguation pages
Ribbon (computing)

From Wikipedia, the free encyclopedia

In GUI-based application software, a ribbon is an interface where a set of toolbars are placed on tabs in a tab bar. Recent releases of some Microsoft applications have embraced this form with an intricate modular ribbon as their main interface.

Ribbons in Microsoft software

This article is written like an advertisement. Please help rewrite this article from a neutral point of view. For blatant advertising that would require a fundamental rewrite to become encyclopedic, use {{ds-stub}} to mark for speedy deletion.

Microsoft originally implemented ribbons as part of its "Fluent User Interface" in Office 2007. The ribbon is formed as a panel that houses the command buttons and icons, organizes commands as a set of tabs, each grouping relevant commands. Each tab has a different set of features which expose the functionality that application offers. For example, while Excel has a tab for the graphing capabilities, Powerpoint does not, instead providing tabs for controlling animation and configuring slide shows. Within each tab, various related options may be grouped together. The Ribbon is designed to make the features of the application more discoverable and accessible with fewer mouse clicks, as compared to the menu-based UI used prior to Office 2007. Moving the mouse scroll wheel while on any of the tabs on the ribbon cycles through the tabs. The Ribbon can be minimized by double clicking the active section's title, such as the Home text in the picture below. Notice the lack of File/Edit menu. The ribbon consolidates the functionality previously found in menus, toolbars and many task panes into one area to increase speed, if you know where commands are. The ribbon UI has also begun to be implemented in other Microsoft software, some applications in Windows 7 such as Paint and WordPad now utilize a ribbon-based UI.

Design guidelines

According to Microsoft their Office 2007 ribbon design guidelines are confidential and an evaluation copy is available when a non-disclosure agreement has been signed.
Eintragsmodell + Prinzipien
Term autonomy

- All terms belonging to one concept should be managed (in one terminological entry) as autonomous (repeatable) blocks of data categories without any preference for a specific term.

- Therefore all terms can be documented with the relevant term-related data categories.

- Term autonomy is necessary for the main term, all synonyms, all variants, and all short forms.

- Term autonomy is not explicitly discussed in theoretical literature.
Concept orientation & term autonomy

TermEntry

Concept

represented by ID-No. and/or classification / notation + Metadata

Language 1  
+ Metadata

Term 1 
+ Metadata

Term 2 
+ Metadata

Language 2  
+ Metadata

Term 1 
+ Metadata

Term 2 
+ Metadata

Term 3 
+ Metadata

Language 3  
+ Metadata

Term 1 
+ Metadata

...
Terminology Metamodel (ISO)
Metamodell in MultiTerm 2009

Metadata

Term level

Concept level

Language level
Metamodel in MultiTerm

**Entry = Concept**

**Language**

**protocol stack**
- Part of Speech: noun
- Status: preferred
- Usage Register: technical
- Context: Another standard architectural model that is often used to describe a network protocol stack is the OSI reference model. This model consists of a seven layer protocol stack.

**stack**
- Part of Speech: noun
- Status: admitted
- Usage Register: colloquial

**stack of interconnected protocols**
- Part of Speech: noun
- Status: admitted

**German**

**Protokollstapel**
- Part of Speech: noun
- Grammatical Gender: masculine
- Context: TCP/IP ist der Protokollstapel, der für die Datenübertragung im Internet hauptsächlich verantwortlich ist und allgemeine Tätigkeiten im Bereich Routing übernimmt (Routing = Vermittlung anhand von IP-Addressen).
Metamodel in TBX

...<text><body>
  <termEntry id='ID0000073578'>
    <descrip type='subjectField'>Materialbeschaffenheit</descrip>
    <langSet lang=de>
      <descrip type='definition'>Maß für die Lichtundurchlässigkeit</descrip>
      <ntig>
        <termGrp>
          <term>Opazität</term>
          <termNote type='partOfSpeech'>Substantiv</termNote>
          <termNote type='grammaticalGender'>f.</termNote>
        </termGrp>
      </ntig>
    </langSet>
  </termEntry>
...<text>
Choice of level

The typology of data categories gives some hints, on which level of the data model the data categories are attached (concept-, language- or term-oriented datCats).

Very often, the level is exactly defined:

- e.g.: *part of speech* at the term-level

But in some cases, the choice of level depends on the objective & philosophy of the termbase:

- e.g.: *definition* at the concept-, language-, term-level
**external collateral ligament**

- **Part of Speech Noun**
- **Status Admitted**
- **Type Full form**
- **Usage register Technical**
- **Grammatical Number Singular**

**Context:** We will discuss reconstruction of external collateral ligament of the knee with tantalum mesh.

**C_Source**

**fibular collateral ligament**

- **Part of Speech Noun**
- **Status Admitted**
- **Grammatical Number Singular**
- **Type Full form**
- **Usage register Technical**

**Context:** Its most distal margin is just proximal to the fibular head where the fibular collateral ligament inserts, and its more proximal aspect is at the superior edge of the anterior arm of the long head of the biceps femoris muscle.

**C_Source**

**lateral collateral ligament**

- **Part of Speech Noun**
- **Status Preferred**
- **Grammatical Number Singular**
- **Type Full form**
- **Usage register Technical**

**Context:** LCL injury is a stretch, partial tear, or complete tear of the lateral collateral ligament (LCL) of the knee. (The term “lateral” means the ligament is on the outside of the knee.)

**C_Source**

**LCL**

- **Part of Speech Noun**
- **Status Preferred**
- **Grammatical Number Singular**
- **Type Abbreviation**
- **Usage register Neutral**

**Context:** The LCL is usually injured by pressure placed on the knee-joint from the inside, which results in stress on the outside of the joint (varus stress).

**C_Source**

---

**definition pertains to all terms in a language.**

(S.E.Wright)
TBX for terminological data modeling

- It reflects best practices (concept orientation and term autonomy)
- It is both an industry standard from ETSI (LISA) and an ISO standard (ISO 30042)
- It builds on existing standards (XML, Unicode, and TMF) [TMF is an ISO standard (16642) for framework for terminology markup languages]
- For new termbase, helps avoid mistakes; for existing termbase, helps assess and improve
- It is being implemented by some tool vendors
Example: MultiTerm 2011 & TBX

- SDL Trados MultiTerm is the market leader software for terminology management since many years with myriads of installations.
- MultiTerm can be used stand-alone, as a client-server application in a multi-user environment, and as a browser-based web TMS.
- MultiTerm allows for / requires user- and application-specific termbase definitions.
- MultiTerm 2011 is the current version of MultiTerm with a lot of new features (termbase modeling is the same as in MT 2007/2009).
Example: MultiTerm 2009 & TBX

MT11 data model
Individual data category definition
Terminological data modeling

Terminological data categories

  - definition, subject field, grammar, context, project code, author, date etc.
  - Data Category Registry (DCR, ISOcat)

Terminological data modeling principles

- ISO 12200 / 16642 / 30042 / 26162
  - meta model, concept orientation, term autonomy, TBX (Termbase eXchange)
Conclusion

- For (computerized) terminology management and termbase design:
  
  - Consult literature and guidelines for terminology management (e.g. Wright/Budin: Handbook of Terminology Management Vol I / II and training course material)
  
  - Follow (ISO) terminology standards (e.g. ISO 704, ISO 1087, ISO 12620, ISO 26162)
  
  - Create guidelines and quality procedures for your own terminology work and your own terminology management solution!
Thank you for your attention

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