Data Modeling and Data Categories for Terminology Management

Sue Ellen Wright
Kent State University
Terminology Summer School 2011
Terminology Management

- There is unfortunately no cure for terminology; you can only hope to manage it.

(Kelly Washbourne)
Overview

- Data modelling principles
  1) Concept orientation
  2) Data categories (datcat)
  3) Term autonomy & datcat repeatability
  4) Elementarity
  5) Dependencies (combinability)
  6) Data modelling variance
  7) Granularity
  8) Shared resources
1) Concept Orientation

- All terminological information belonging to one concept, including all terms in all languages and all term-related and administrative data are be stored (or displayed) in one terminological entry.

- 1 concept = 1 (real or virtual) terminological entry.
Concept Orientation

- Modern Terminology Management Systems (TMS)
  - Follow the concept-oriented approach
  - Support features for consistent concept entries.
  - Discourage doublettes (“double entries”) for the same concept
  - Allow for homographs (different concepts represented by the same term)
2) Data Categories

Prior to computerization, data categories were not discussed in detail in terminology theory.

Initial approaches to data categories involved describing “fields” in paper forms for recording terminological data offline.
<table>
<thead>
<tr>
<th>Fachgebiet(e)</th>
<th>Subject field</th>
<th>Sprache</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teilbestand</td>
<td>Projektcode</td>
<td>Notation</td>
<td>Klassifikations-schlüssel</td>
</tr>
<tr>
<td>Subset</td>
<td>Project code</td>
<td>Notation</td>
<td>Classification code</td>
</tr>
<tr>
<td>Benennung</td>
<td>Term</td>
<td>Quelle</td>
<td>Source</td>
</tr>
<tr>
<td>(Kurzformen, Abkürzungen, orthographische Varianten)</td>
<td></td>
<td>Grammatische Angaben</td>
<td>Grammar</td>
</tr>
<tr>
<td>Definition(en)</td>
<td>Definition</td>
<td>Quelle(n)</td>
<td>Source</td>
</tr>
<tr>
<td>Kontext(e)</td>
<td>Contexts</td>
<td>Quelle(n)</td>
<td>Source</td>
</tr>
<tr>
<td>Bemerkungen</td>
<td>Comments (note)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonyme (falls nicht als separater Eintrag, dann mit Angabe der Quelle)</td>
<td>Synonyms</td>
<td>Quelle(n)</td>
<td>Source</td>
</tr>
<tr>
<td>Erfasser – Datum</td>
<td>Editor/Date</td>
<td>Bearbeiter - Datum</td>
<td>Entry Class</td>
</tr>
</tbody>
</table>
Packaging Data

- Model for visualizing information
  - An amorphous flow of undifferentiated “stuff”
  - An aggregate of individual elements (little packages, components) that can be identified, delimited, organized (modeled), stored, retrieved, manipulated, and reused
  - Stuff people can figure out if they think about it
  - Stuff a computer program can automatically recognize and process

2) Data Categories
The Shopping Basket Model

- A shopping basket is a container to put stuff in.
- The stuff in the basket *can be* tossed in loose without packaging.
- Stuff gets lost.
- Stuff gets mixed up.

2) Data Categories
Bulk Product Loading

Unpackaged rat poison

Loose flour

Unwrapped foodstuffs

2) Data Categories
Packaging Stuff

- Keeps products clean and uncontaminated
- Makes them easy to identify
- Makes them easy to store
- Makes them easy to reuse
Main Entry Term shorter form of a long name, such as an email address, a directory, or a command Fahey:7

Actually, all text-type addresses, long or short, are aliases of the IP address which is numerical only. Fahey:7

Aliases in real-time chat are usually referred to as nicknames and handles nickname noun Synonym

When you start an IRC session, you specify a nickname, up to nine characters, which you can also change at any time. Osborne94: 413 handle noun

The nickname you assign yourself when conversing in discussion groups. Falcón: 94 Use only in colloquial situations. alias noun m un nombre corto o apodo que se utiliza para enviar mensajes a aquellas direcciones de uso más frecuente. Carballar94:122 Un alias representa a un grupo de personas que se envía un correo a un alias, todo el grupo de personas que lo forman. Carballar94: 112

2) Data Categories
2) Data Categories

acetabulum  
ACL  
acute case  
acute repair  
adductor tubercle  
angiogram  
angiography  
anterior cruciate ligament  
arteriography  
bouton en chemise  
dislocation  
drop foot  
televator mechanism  
external collateral ligament  
external sciatic popliteal  
fibular collateral ligament  
foot drop

**Organized Data**

---

**Entry number: 61**

**Subject:**

**Graphic:**

---

**G_Source:**

http://www.zimmer.co.za/web/images/patient_education/colligears01.jpg

---

**EN English**

Definition: A strong, rounded, fibrous cord, attached, above, to the back part of the lateral condyle of the femur, immediately above the groove for the tendon of the Popliteus; below, to the lateral side of the head of the fibula, in front of the styloid process.

**D_Source:**

http://en.wikipedia.org/wiki/Fibular_collateral_ligament

**Note:** The ligament connects the lateral epicondyle of the femur with the lateral side of the head of the fibula and that helps to stabilize the knee by preventing lateral dislocation.

**N_Source:**

http://www2.merriam-webster.com/cgi-bin/mwmednlt?book=Medical&xv=lateral%20collateral%20ligament

---

**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

---

**external collateral ligament**

---

---
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
Specifying the Data Category

- `/term/` is a data category
- `/xml:lang/` (xml language representation) is a data category
- We specify what `/term/` is with a definition: verbal designation of a general concept in a specific subject field
- We can also constrain the `/term/` category: terms are reported as *plain text* and not as graphics or numbers (although a number can be used as a term).
- We use the word *constrain* to mean that we limit what the something can do.
Kinds of Data Categories

- complex data category (MultiTerm: index, text, picklist)
  - data category that has content of some sort

- open data category (MultiTerm: index, text)
  - complex data category whose conceptual domain is not restricted to a set of values (non-enumerated CD)

- closed data category (MultiTerm: picklist)
  - complex data category whose conceptual domain is restricted to a set of identified simple data categories making up its value domain (enumerated conceptual domain)

- simple data category (MultiTerm: picklist value)
  - data category with no conceptual domain; a member of a value domain
Data Categories

- **open data category**
  - **complex data category** whose content (conceptual domain) is not restricted to a set of values (non-enumerated conceptual domain)
  - Example: /term/, /definition/, /note/
  - Open data categories are still constrained by their specifications and definitions.
    - Only a term should go in a term field.
    - Only a definition should go in a definition field
    - But no one can say precisely what terms or definitions there are
Sample Open Data Category

- Data categories can be combined:
  - en:term = a term in English
  - An instance of a data category is a data element.

- en:term = external collateral ligament is an individual data element.

- en:term = fibular collateral ligament is an another individual data element.
Picklists: Closed Data Categories

- \textit{grammatical gender} is a data category
- \textit{grammatical gender} has defined values:
  - masculine
  - feminine
  - neuter
  - other
- m/f/n/o are also called picklist values, permissible values or permissible instances.
Permissible Values

- Together, permissible values comprise a value domain.
- No other values are allowed unless a change is made in the system.
- Terminologists sometimes call these values *simple data categories*. They are values, but they do not themselves have dependent values.
- Each has its own value meaning.
Data-Element Related Features

- Data Element Autonomy (repeatability)
  - Elementarity of data elements
  - Dependency (combinability)

- Data Modeling Variance
  - Elementarity
  - Granularity

- Shared Resources

- Metamodel issues
  - Standards
Constrained Data Categories

- Some data categories can hold only certain kinds of data.
  - Example: We can create a data category called *Graphic* and set the data type of that datcat to *Media*
  - Example: We can create a data category called *Date* and set the data type of the datcat to *Date*
Term Autonomy & Repeatability

- All terms should be managed as autonomous (repeatable) blocks of data categories without any preference for a specific term
  - Each term documented with the relevant term-related data categories
  - Term autonomy applicable for preferred term, synonyms, variants, & short forms
  - Term autonomy not explicitly discussed in other theoretical literature
Term Autonomy

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

Language 1
Term 1 + DatCatSet
Term 2 + DatCatSet

Language 2
Term 1 + DatCatSet
Term 2 + DatCatSet
Term 3 + DatCatSet

Language 3
Term 1 + DatCatSet

3) Autonomy
Here each term has its own DatCatSet.

**air bag**

**Gebrauch norm**

Quelle ISO/DIS 12097-1

Definition A flexible material, forming an enclosed volume that receives the gas from the inflator and restrains the occupant.

D-Quelle ISO/DIS 12097-1

Kontext The actual air bag in an air bag restraint system; typically made of a polyamide fabric, often coated with neoprene on the inside; newer air bags have no coating to facilitate recycling.

K-Quelle Schmitt.1992

**airbag**

**Benennungstyp Variante**

Quelle Wyhlidal.1994

Kontext The headform shall strike the test component at a speed of 24.1 km/h or, in the case of components which cover an uninflated airbag, at a speed of 19.3 km/h; this speed shall be achieved either by the mere energy of propulsion or by using an additional impelling device.

K-Quelle ECE-R 21, Annex IV, 1.4.2

**bag**

**Benennungstyp Kurzform**

Quelle Schmitt.1992

Kontext Porosity is characteristic of the bag fabric measurable as gas permeability.

K-Quelle ISO/DIS 12097-1

**air cushion**

**Regionalcode AE**

Quelle Schmitt.1992

Kontext Upon vehicle collision a pyrotechnic gas generator inflates an air cushion (airbag)(...).

K-Quelle Bosch/VDI.1993, S. 718

**air bag cushion**

**Kunde BWM**

Quelle Schmitt.1992

---

3) Autonomy
Here terms are hidden in /variant/ and /synonym/ fields.

**3) Autonomy**
Embedded or Hidden Terms

- **en:term=** external collateral ligament
- **definition=** ligament that bla bla bla …
- **note=** bla, bla, bla, also sometimes called the “fibular collateral ligament”.

**Problem:**
- You can’t look up the buried term.
- It doesn’t become part of any concept system.
- It ends up in a /note/ element instead of a /term/ element.
Here the synonyms and variants are strung together in one /term/ field & we can’t document info about individual terms.

**air bag, airbag, bag, air cushion, air bag cushion**

**Quelle**  ISO/DIS 12097-1

**Definition**  A flexible material, forming an enclosed volume that receives the gas from the inflator and restrains the occupant.

D-**Quelle**  ISO/DIS 12097-1

**Kontext**  The actual air bag in an air bag restraint system; typically made of a polyamide fabric, often coated with neoprene on the inside; newer air bags have no coating to facilitate recycling.

**K-**Quelle  Schmitt.1992

**Luftsack, Luftkissen, Prallkissen, Prallsack, Sack, Airbag Genus m.**

**Quelle**  Schmitt.1992

**Definition**  Eigentlicher Luft sack in einem Airbag Rückhaltesystem; meist ein Gewebe aus Polyamid (Nylon), das entweder innen beschichtet ist (Neopren) oder zwecks besserer Recyclingfähigkeit (Sortenreinheit) unbeschichtet bleibt.

D-**Quelle**  Schmitt.1992


**K-Quelle**  Ostertag.1996, S. 67

**Notation**  1-1-2-2-2-3-1
Data Element Elementarity

- 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
Common Elementarity Errors (1)

- 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/.
Common Elementarity Errors (2)

- 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.
Common Elementarity Errors (2)

- **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.
Modelling Dependencies (Combinability)

- Dependencies between data categories
- ISO 12620:1999 provides a “simple hierarchy” of data categories
  - $grammar = \text{term-related}$: 
    - $grammar$ is dependent upon $term$
- Other dependencies
  - $source$ is dependent upon $definition$
  - for additional definitions, additional sources are needed
  - Individual source indicated for different items (term, definition, context, etc.)
Data Modelling Variance

- 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
Modelling 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

6) Variance
Granularity

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

- Low granularity:
  - Ex: /grammar/ m,n,s (masculine noun singular)

- High granularity
  - /part of speech/ = noun
  - /grammatical gender/ = masculine
  - /number/ = singular
Granularity

- Advantage of granularity: retrievability
- Disadvantage: more work
- Best practices:
  - General agreements on what constitutes an acceptable level of effort in order to achieve a desired level of granularity makes it easier to share data.
  - How can you automatically convert low granularity to high granularity?
# Minimum Granularity

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>John Doe</td>
</tr>
<tr>
<td>Address</td>
<td>267 Prospect Street Kent, Ohio 44240</td>
</tr>
</tbody>
</table>
Increased Granularity

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td>Doe</td>
</tr>
<tr>
<td>First name</td>
<td>John</td>
</tr>
<tr>
<td>Street address</td>
<td>345 North Prospect St.</td>
</tr>
<tr>
<td>City</td>
<td>Kent</td>
</tr>
<tr>
<td>State</td>
<td>Ohio</td>
</tr>
<tr>
<td>Zip Code</td>
<td>44240</td>
</tr>
</tbody>
</table>
Power of Granularity

- *zip code* (postal code) can be searched and manipulated
- Extra effort in entering a separate field yields search power later, and in this case functionality.
- Sorting by postal code makes mail delivery more efficient & saves money.
- Know ahead of time what you want to do with your data when deciding on granularity.
Elemental Nature of Data Elements (Violation)

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td>Doe</td>
</tr>
<tr>
<td>First name</td>
<td>John and Jane</td>
</tr>
<tr>
<td>Street address</td>
<td>345 North Prospect St.</td>
</tr>
<tr>
<td>City</td>
<td>Kent</td>
</tr>
<tr>
<td>State</td>
<td>Ohio</td>
</tr>
<tr>
<td>Zip Code</td>
<td>44240</td>
</tr>
</tbody>
</table>

7) Granularity
# Shared Resources

<table>
<thead>
<tr>
<th>Index</th>
<th>Oxford1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor1</td>
<td>Fowler, H.W.</td>
</tr>
<tr>
<td>Editor2</td>
<td>Fowler, F.G.</td>
</tr>
<tr>
<td>Editor3</td>
<td>Allen, R.E.</td>
</tr>
<tr>
<td>Location</td>
<td>Oxford</td>
</tr>
<tr>
<td>Press</td>
<td>Clarendon</td>
</tr>
<tr>
<td>Year</td>
<td>1990</td>
</tr>
</tbody>
</table>

Entry 1
Source ID: Oxford1990

Entry 2
Source ID: Oxford1990

8) Shared
Shared Resources

- Graphics
- Charts
- Audio
- Video
- Drawings
- Disk Archives
- Responsibility Records
Advantages of Shared Resources

- Data category configured as a link or with a link
- Shared items stored only once
- Multiple references linked to same resource
- Change or update only one file or entry
- File formats (images, audio, etc.) preserved externally and objects linked
- Terminology files of manageable in size
clutch slave cylinder
Concept Entry Level

- Concept level
  - Subject field(s)
  - Administrative information
  - Graphic (option)
  - Definition & related info (option)
  - Note (repeatable anywhere)

9) Metamodel
Language Level

- Object language info
- Graphic (option) & related info
- Definition (option) & related info
- Administrative info
- Note (repeatable anywhere)
Term Information Group Level

- Term
- Term-related info
- Descriptive info
  - Definitions, other concept description
  - Graphic
- Administrative info
Term-Related DatCats

- Term
  - Part of speech
  - Grammatical gender
  - Grammatical number (use when necessary)
    - Plural form
  - Term type (Type) (see next slide)
  - Status
  - Regional label
  - Pronunciation
  - Register (usage register)

9) Metamodel + “Vocabulary”
Term Type

- Main entry term
- Synonym
- Abbreviation
- Full form
- Variant
- Phrase
- Collocation

9) Metamod + “Vocabulary”
9) Metamodel + “Vocabulary”
9) Metamodel + “Vocabulary”
9) Metamodel + “Vocabulary”
Definition-Related Info

- **Definition**
  - Source ID
    - (Points to shared resources)
  - Definition type
    - Translation?

- **Choice of levels**
  - Concept (one def for all languages)
  - Language (one def for each language)
  - Term (one def for each term in descriptive work)
Context-Related Info

- Context
  - Source ID
    - (Points to shared resources)
  - Context type
    - Translation?
  - Both term and concept-related
  - Always associated at the term-level, never at lang or concept level

9) Metamodel + “Vocabulary”
Administrative Information

- Administrative info
  - Responsibility
  - Date
  - Authorization
  - Role
  - Various sorting subsets
    - Business unit
    - Customer
    - Computing environment
    - Product class, etc.
  - Applicable at multiple levels
Other Data Categories

- Concept relations
- Notes
- Other administrative information
- Bibliographical information
- Special categories, e.g.:
  - For standardization
  - For inventory control
Data Category Types

- Graphics (link to multimedia file)
- Terms
  - By language
  - Synonyms
  - Term-related information
- Text fields
  - Definitions, Contexts, Notes, etc.
- Automatically inserted administrative data
- Links
MultiTerm Levels & Types

- **Index**
  - Terms
  - Term-like elements
    - (collocations, boilerplate)
- **Text** (definitions, contexts, notes, etc.)
  - Free-form data entry
- **Picklist**
- **Multimedia file**
- **Number**
- **Date**
Graphics can be used as resources for the whole entry, as a reference for a single language, or even to illustrate a single term. On my Christmas list: I want this kind of graphic to be an image map where I can click to the various terms cited on the image. Note the inclusion of the source of the graphic.

EN English

Definition A strong, rounded, fibrous cord, attached, above, to the back part of the lateral condyle of the femur, immediately above the groove for the tendon of the Popliteus, lateral side of the head of the fibula, in front of the styloid process.

D_Source http://en.wikipedia.org/wiki/Fibular_collateral_ligament

Note The ligament connects the lateral epicondyle of the femur with the lateral side of the head of the fibula and that helps to stabilize the knee by preventing lateral dislocation.

N_Source http://www2.merriam-webster.com/cgi-bin/mwmednolm?book=Medical&va=lateral%20collateral%20ligament

external collateral ligament

10) MultiTerm Meta-Categories
In this entry: /definition/ pertains to all terms in a language.

Synonyms appear together

Linguistic info, term type & status marked

Contexts associated with terms

Sources indicated for definitions, contexts, notes
Graphic anchored at the level of the Spanish language shows Spanish terms.

10) MultiTerm Meta-Categories

**ligamento colateral externo**
- Part of Speech: Noun
- Status: Preferred
- Grammatical Gender: Masculine
- Grammatical Number: Singular
- Type: Full form
- Usage register: Technical
- Context: Ligamento colateral externo: Se origina en una depresión ósea 1.4 mm proximal y 3.1 mm posterior al epicóndilo externo y se inserta 8.2 mm posterior a la cabeza fibular anterior y 28.4 mm distal al extremo del proceso estiloides.

G. Source: [http://www.umm.edu/esp_imagepages/8841.htm](http://www.umm.edu/esp_imagepages/8841.htm)
D. Source: [http://www.umm.edu/esp_imagepages/8841.htm](http://www.umm.edu/esp_imagepages/8841.htm)

LCE
- Part of Speech: Noun
- Related Term: L.C.E.
Accessing Standard DCs

http://www.isocat.org