TERMINOLOGY STANDARDIZATION AND TRANSLATION STANDARDS

Christian Galinski
Infoterm & ISO TC 37 Secretariat

Reinhard Weissinger
ISO Central Secretariat

TKE 2010
Dublin 2010-08-14
Terminology standardization

- Standardization of terminologies
- Standardization of terminological principles and methods
- Standardization of methodological and technical aspects of terminology applications
- Related standards of all sorts
- ISO/CDB – Concept DataBase

Translators should know about
- standardization in general (a bit)
- pertinent standardization and certification
- standard-compliant technology (quite a bit)
Terminology is knowledge representation

- Terminology as a type of language resource is a set of terms that represent concepts of a specific knowledge domain

- Terminologies also include non-linguistic concept representations such as graphical symbols, formulae, etc.

- Terminologies are compiled according to terminographical methods and presented as dictionaries, databases, etc.
Terminologies

- are means of domain communication
  play a crucial role in education and all situations of professional and scholarly communication (incl. translation and interpretation)

- are means of access to other kinds of information (objects)
  indispensable for information/content management, archiving, etc.
Terminology standardization

- **Standardization of terminologies**
  - Terminological data
    - Linguistic and non-linguistic representations
      - **Designation(s):** terms, abbreviations, graphic symbols, formulas, acoustic symbols, etc.
      - **Description(s):** definitions, explanations, non-linguistic [descriptive] representations, etc.
    - **Source-related data** & copyright info
    - Data management related data (field, record, holding)
    - Classification (multiple)
  - Terminology-related data: names, phraseology, ...

- **Standardization of terminological principles and methods**

  ➔ generic for many types of other items of structured content
Complex content items

Increasingly terminological information and other kinds of **structured content**
(at the level of lexical semantics)

- are combined with each other
- embedded in each other

often forming **complex content items**
(still at the level of lexical semantics)
Structured content: Traffic informatics

Way to the airport – turn right in 5 km

Way to the train station – down to the right

ZONE = verbal
red ring = (morphology) prohibition sign
30 = micro-proposition: max speed 30km/h

→ variable message sign boards
communicating with car-driver system

TKE-WS, Dublin 2010-18-14
Structured content: Product catalogues

e.g. complex entry in a product catalogue

- **Name of company** (® enterprise)
- **Name of product** (model) (™ enterprise)
- **Generic name of product** (e.g. © HS)
- **Class** (name under which the product falls) (e.g. © eCl@ss)
- **Verbal/textual description** (© enterprise)
- **Picture** (© enterprise)
- **Technical data**
  - (unified) branch properties (e.g. © OAGi)
  - Standardized characteristics (e.g. © DIN)
  - Enterprise product specific data (e.g. for collaborative business)
  - Enterprise internal data (maybe confidential/secret)

225/55/16 V
Structured content: Kanji Flashcard 休

8. **Meaning of look-alike.** The English meaning of the look-alike.

9. **Radical.** The main kanji radical. Radicals are traditionally used to order entries in Chinese and Japanese character dictionaries.

10. キュウ
11. やす・む/やす・まる/やす・める
12. rest, take a day off, relax
13. person tree
14. 1. ていきゅうび a regular holiday (for a store) 2. ひとやすみする to take [have] a (short) rest 3. やすまる to be [feel] rested; to be relieved 4. やすむ to rest; to take a day off; to sleep 5. やすみ (a) rest; a holiday; (a) vacation; a day off 6. きゅうじつ a day off; a holiday

1. **Main kanji.** The main kanji character is written in a large, brush stroke typeface.
2. **Kanji compounds.** Six compounds containing the main kanji

TKE-WS, Dublin 2010-18-14
ISO/TC 37 (3) – from ISO/TC 37 BP

Language resources

- Standardization is also needed for other language resources (mono- and multilingual), e.g. speech data, written (full) text corpora, lexical (general language) corpora and their processing methods

- Relevant research areas are computational linguistics and computational lexicography, language engineering, etc., which have provided industrial best practices to be turned into official standards

- This process will contribute to the further development of the language industries at large

- Similar to terminologies, language (and other content) resources in general have to be considered as multilingual, multimodal and multimedia from the outset
Terminology management + language and content resource management

- **Language resources:**
  - Text corpora → **tagging** (on the basis of grammar models)
  - Lexicographical data
    - Words
    - Collocations
    - Morphology
  - Terminology & terminological phraseology
  - Speech data

- **LR management:**
  - Preparation, maintenance, exchange, ...
  - Metadata (incl. bundling/bindings etc.)
  - Data modelling & metamodell(s)
  - Data exchange / interoperability
  - etc.
Structured content today (1)

- Terminology
  - Nomenclature, taxonomy, typology, ...
  - Glossary, vocabulary, ...
  - Terminological phraseology
  - Graphical symbols and other non-linguistic representations?
  - Properties, characteristics, attributes, ...
  - Ontologies
    - Names, names, names, ...

- Thesauri, classification schemes, keywords

- Encyclopedic (knowledge) entries
  - Knowledge-enriched terminology entries
  - (explained) proper names, ...

- Ontologies, topic maps, ...

- Lexicographical data and other language resources
- Other content resources

are often contradictory, NOT coherent, integrated, reliable, ...
Structured content today (2)

- According to content management (technical p-o-v):
  - Texts: → translation, localization, internationalization...
  - Speech: → communication...
  - Image: → CAD/CAM...
  - Multimedia: → video, presentations...

- At the level of lexical semantics (content p-o-v):
  - Terminology → basis of domain knowledge
  - Language resources
  - Other content resources
    - incl. non-verbal representations
  - Meta-content – i.e. content about content
  - Metadata – i.e. data about data (data categories)
Practical problems

Standardization of terminologies (and other kinds of structured content)

- is difficult (if done properly)
  needs a minimum of methodological experience
- lacks “customized” training
  for terminology standardizing experts in given subject fields
- is not “attractive” (for up-and-coming experts)
  hence the common attitude: “leave it to the old hands”
- is time consuming and therefore expensive
  not “rewarding” enough (compared to the great efforts involved)
- no “user-friendly” system support
  for cooperative committee work in terminology standardization
- little coordination
  across all standardization (incl. terminology standardization)
- only a limited degree of acceptance
  for standardized terminologies beyond standardization
- “traditional” working method out-dated, etc.
  e.g. working group or sub-committee of terminology experts
STRUCTURED CONTENT DEVELOPMENT

- Time consuming ➔ costs
- Cost of preparation? calculatable, but… ➔ maintenance: quality, reliability, liability, …
- Traditional methods ➔ web-based methods
- Duplication of efforts? ➔ content management
- Application of tools ➔ technical interoperability
- Multilinguality ➔ localization principles
- Distributed work ➔ workflow management
- eContent ➔ mContent

➔ STANDARDIZATION ➔ INTEROPERABILITY
Content interoperability standards

- Content-related requirements
- Workflow methodology
- Metadata and metadata repositories
- Data modelling principles and requirements
- Micro datamodels
- Metamodels
- Content repositories
- Federation of repositories
- Business models (incl. copyright management...)
- ...

TKE-WS, Dublin 2010-18-14
Terminology standardization

- Hitherto prominently verbal-linguistic term-oriented approach to terminological data management needs to take non-linguistic representations of concepts as fully equivalent to verbal-linguistic representations into account.

- Thus a generic datamodel can be achieved, which is applicable to all kinds of “structured content” (here: content items at the level of concepts or lexical semantics).

- Since terminological data and other kinds of structured content have a lot in common, it seems appropriate to handle them with one and the same theoretic-methodological approach.
Increasing use of DBs in standards development

Result of a survey done by ISO Central Secretariat:

- June 2005: Approximately 15 TCs/SCs
- June 2007: Over 40 TCs/SCs
- March 2006: First DIS disseminated in the form of a database (ISO/TC 61/SC 1 – *Plastics vocabulary*)

→ Emerging approach to standards development
ISO/CDB – Content (1)

ISO standards containing terminology

- Vocabulary standards: app. 800
- Other standards with terminology: app. 8000
- App. 180,000 ~ 200,000 terms in ISO standards
- → Cooperation with EAFTerm (China, Japan, Korea):
  ISO/CS obtained around 120,000 terminology records from ISO standards through a project implemented by CNIS
ISO/CDB – Content (2)

- **Graphical symbols**
  - App. 4500 (in ISO standards)
  - App. 1000 (in IEC standards)

- **Other types of representations (codes, product properties etc.)**
  - Numbers - ?????
Applicability of the DB approach

Standard consists in full or in part of a "collection of items", e.g.

- graphical symbols
- terms and definitions
- product properties
- data dictionaries of all types
- classification systems
- codes (for various types of objects)
etc.
Current use of databases by ISO committees

Advantages:
- Large number of items can be easier managed and maintained
- Consistency can be easier ensured
- Database tools readily available

Challenges/problems:
- Lack of support in terms of IT infrastructure
- Grass root developments
- Lack of compatibility (e.g. regarding data categories), maintenance problems
- Access modalities to content are unclear
TMB AHG (1)  
“Standards as databases”

Established by ISO Technical Management Board in June 2005 with the mandate to investigate issues relative to standards as databases, incl.

- Possibilities for harmonization
- Implications related to standards development procedures & technical infrastructure
- Access modalities, incl. commercial aspects
TMB AHG (2) – Main output

- Development and maintenance procedure for standards in database format
  (approved by TMB in June 2007 – Annex ST to ISO Supplement to the Directives)

- Specification of basic functions of a concept database

- Data model for CDB (extension of ISO 16642:2003 to address all kinds of concepts irrespective of the type of their representation)

- Definition of four access layers related to different types of objects in the CDB (approved by CPSG in June 2007)
Use of databases

Use type 1: For maintenance *after* development

Approved Work Item \(\rightarrow\) traditional development process \(\rightarrow\) Publication of a standard as a document \(\rightarrow\) extraction \(\rightarrow\) upload to Database

Use type 2: For development from scratch

Approved Work Item \(\rightarrow\) storage of proposed/draft concepts \(\rightarrow\) Database \(\rightarrow\) development \(\rightarrow\) concept status set to « approved », « standard » \(\rightarrow\) Database

TKE-WS, Dublin 2010-18-14
ISO/CDB - ISO procedure (1)
(Annex ST to ISO Supplement)
ISO/CDB – Procedure (2)

- "Source/Master" of the standard:
  - Stored in the database (not a document!)

- Maintenance and development of standards by a committee through defined processes

- Procedure includes four sub-processes:
  
  - (1) Development of new standards (start with New WI Proposal)
  
  - (2) Maintenance of existing standards (start with Change Request)
  
  - (3) Withdrawal process (start with Change Request)
  
  - (4) Systematic review of standards
ISO/CDB – Procedure (3)
« Normal » DB-procedure [2 – 6 months]
Implementation

- Cooperation agreement with software partner
- Cooperation mode and input from ISO committees to be organized
- Implementation time frame: 2007 Q4 – 2009 Q2 (first release)
- Review of implications of the new procedure & implementation by ISO/CS
- Integration with the production chain in the Central Secretariat
ISO/CDB - Access modalities

- Free access-layer (thumbnails, basic record elements) for navigation in the content
- Terminology freely available ("ISO electronic dictionary")
- Web-access to individual items
- Federated development/maintenance of content on the basis of role based permissions
  - Project teams, PLs, PEs / Review groups / Other committees

http://cdb.iso.org
Thank you for your attention

Infoterm – International Information Centre for Terminology

Gymnasiumstrasse 50
1190 Vienna – Austria

Tel: +43-1-4277-58026
Fax:+43-1-5876990
infopoint@infoterm.org
http://www.infoterm.info

ISO/TC 37 “Terminology and other language and content resources”
ISO/CDB – Data model

- Extended data model based on ISO 16642:2003
  - Concept-oriented
  - Support for concept representations other than terms (graphical symbols, codes etc.)
  - Multilinguality

- Model needs to be further tested with other types of representations
Structured content (1): content entities at the level of lexical semantics

Types of representations of meaning/knowledge:
- Visual
- Audio
- Tactile/haptic
- Smell/taste
- Other

Non-verbal

Verbal

Non-verbal

Audiovisuell / video

Multimedia

Future semantics \rightarrow ITU-T X.1081 “Telebiometrics MMF”
TYPES OF SD-RELATED REPOSITORIES

- Classification etc.
  → multiple
- Properties
  → acc. to types
- Data dictionaries
  → acc. to types
- Metadata
  → acc. to types
- Terminologies
  → acc. to types & domains
- Ontologies
  → acc. to types & domains

→ huge amounts of repository items
to be taken care of in federated registries
Terminological knowledge engineering

- **Terminology science** provides the methodology for the preparation, recording and processing (as well as re-use) of terminological data.

- **Terminography** supplies the tools for the efficient preparation and processing of terminological data which in turn are further processed into dictionaries, vocabularies, terminology databases, etc.

- **Terminological knowledge engineering** provides the tools to represent, manage and access knowledge of different degrees of complexity.

- Knowledge(?)/content management cannot be efficient without a strong terminology component (comprising terminological data, methods and tools).
DATA MODELS & METADATA

1. ISO/TC 37 → fundamentals of multilingual mContent development (incl. language resources and LR management)

2. application areas: eBusiness, eHealth, eLearing, eGovernment, product data management, ...

3. other initiatives and consortia:
   OIDDI - Open and Interoperable Data Dictionaries Initiative
   OASIS, OMG, OAGi, ...

→ (informal) coordination & harmonization: MoU/MG

TKE-WS, Dublin 2010-18-14