

Content Interoperability Standardization and harmonization

Christian Galinski

Infoterm & ISO TC 37 Secretariat

OASIS 1st Int'l Workshop:
**"Technology for independent living
an OASIS beyond 50"**
Brussels 2008-10-10

Overview

- Background
- eContent → mContent
- **Structured content** at the level of lexical semantics
The challenge → **content integration**
- Kinds of structured content
- **Accessibility – Distribution - Development**
(incl. maintenance & archiving)
- **Federation**
- **Consequences for theory and methodology**

Personal background

- **Japanese studies & „communication“ studies**
applied linguistics, information & documentation, social anthropology, Japanese and Chinese...
- **1986: Director of Infoterm & Secretary of ISO/TC 37**
→ *standardization of terminological principles and methods*
→ *extended towards language and content resources*
- **language: important part of communication -**
BUT: communication is broader than language
- **efficient communication** to a large degree:
is based on good structured content
- **for critical communication:**
standardization and harmonization is necessary

Professional experience since 1970s

- **Difficulties with language resources:**
 - Terminology, BUT „phraseology“?
 - Lexicographical data, BUT collocations?
 - Abbreviations? \leftrightarrow proper names
 - Proper names?
 - Classification schemes & thesauri
 - Historical / legacy data ...
- **Access to other content resources:**
 - Letter symbols and other symbols
 - Non-verbal representations
 - All kinds of resources... \rightarrow content integration?

Terminology & content integration III

- 1990s: „language“ is becoming an issue at political level and in industry;
- emergence of localization (L10N) and internationalization (I18N) in the wake of globalization (G10N)
- LISA – Language Industry Standardization Association;
- terminology (application) systems;
- Google;
- ELRA (1995) – European Language Resource Association;
- TermNet (1988);
- IITF (1989) ...

Content interoperability (ContI Op)

Terminology & content integration IV

After 2000:

„language“ is fully established as a **strategic issue** at political level and in big industry;

- LRM – language resource management (e.g. as SC 4 in ISO/TC 37);
- UNESCO: Guidelines for terminology policies;
- EuroTermBank: new business models;
- ISO/CDB: Concept DataBase;
- LISA: TM – terminology management;
- **Wikipedia** → Encyclopaedia Britannica etc. online;
- **Semantic Web** (Tim B. Lee) → semantic web ...

What is content?

- **Digital content: eContent → mContent**
+metadata, +meta-content, +broadband →ubiquity and pervasiveness
- **eContent NOT only content industries**
 - **Structured content** as means of communication infrastructure
 - **Structured content as product** and services
 - incl. public content
- **here: structured content** at the level of lexical semantics
 - Linguistic and non-linguistic content items
 - Content for knowledge representation
 - Content for domain communication
 - Content as product: +DRM and business models

eContent → mContent

Structured content at the level of lexical semantics

■ **New aspects need to be considered**

- Concept representations can be verbal and non-verbal
→ representation autonomy
- ICT technology and new applications have an impact on the accessibility, distribution and creation of structured content
- Economic factors: new business models
- + new findings in brain research
- software development still very deficient!

■ **Theories and methodologies of several domains dealing with structured content need to be aligned, in order to become interoperable**

New world: mContent

Convergence of mobile communication and mobile computing

→ MCC (mobile computing and communication)

- Wired / wireless / satellite → **no difference**
→ increasingly the end user is permanently mobile
- Ubiquity + Pervasiveness + Broadband
- Natural language user interface:
written ↔ spoken → **convergence MCC** + more modalities
- Increasingly 'semantic' communication $m \leftrightarrow m$
→ BUT: user wants USE → **NOT technology**

Software development in general is still poor with respect to requirements of content integration and interoperability!

eContent → mContent: **ONE methodology**
at least for structured content

+ CONTENT DEVELOPMENT

→ web-based, distributed, cooperative creation of structured content

➤ **multilingual**

➤ **multimodal**

➤ **multimedia**

complying with

➤ **multi-channel** output

➤ **eInclusion** requirements

Re-use in applications:

- eLearning
- eGovernment
- eHealth
- eBusiness
- other e...s

→ **Co-operation**

→ **Interoperability**

→ **METHODOLOGY STANDARIZATION**

KINDS of STRUCTURED CONTENT

- **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+ } **incl.**
 - Language resources+ } **non-verbal**
 - Other content resources } **representations**
 - **Meta-content** – i.e. content about content
 - **Metadata** – i.e. data about data (data categories)



⚓ Consultation of the International Hydrographic Dictionary ⚓

Preliminary version (05/05/2000, reduced database) [Jean-Luc Husson](#)

Select from fields below, enter your search string and then press the SEARCH button

French

Term

Contains

courant

Case



[Help](#)



courant côtier [HR-1074] - n

Définition :

Courant de direction générale parallèle à la côte. Toutefois cette définition ne s'applique qu'au courant que l'on rencontre à l'extérieur de la zone de déferlement lorsqu'une telle zone existe.

coastal current [HR-1074] - n

Definition :

A relatively uniform drift usually flowing parallel to the shore in the deeper water adjacent to the surf zone.

corriente costera [HR-1074] - n

Definición :

Corriente de deriva relativamente uniforme, por lo general paralela a la costa, en las aguas más profundas adyacentes a la zona de rompientes.

courant de densité [HR-1075] - n

Définition :

Courant provoqué par le gradient horizontal de densité de l'eau.

Voir aussi :

HR-1094

density current [HR-1075] - n

Definition :

A gradient current caused by the horizontal gradient of water density.

See also :

▪ HR-1094

corriente de densidad [HR-1075] - n

Definición :

Corriente de gradiente causada por el gradiente horizontal de densidad del agua.

Ver también :

▪ HR-1094

courant de dérive [HR-1077] - n

Définition :

Courant océanique superficiel de faible vitesse et

drift [HR-1077] - n

Definition :

A wide, slow-moving current principally caused by

velocidad de deriva [HR-1077] - n

corriente de deriva - n

aporte - n

deriva - n

STRUCTURED CONTENT: Example 1: 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

STRUCTURED CONTENT

Example 2: product catalogues



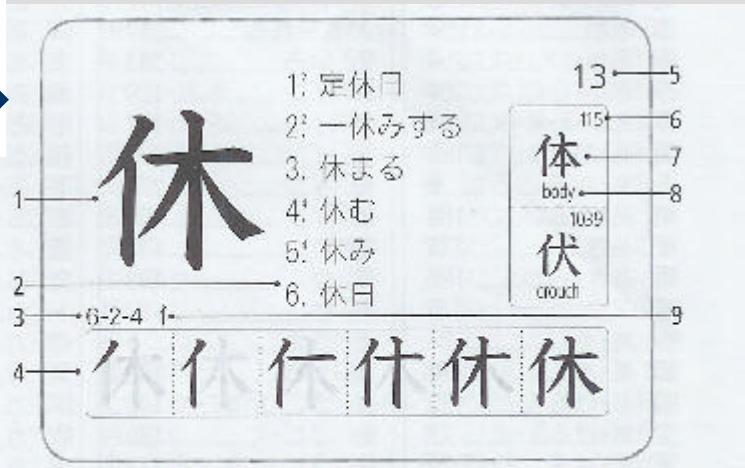
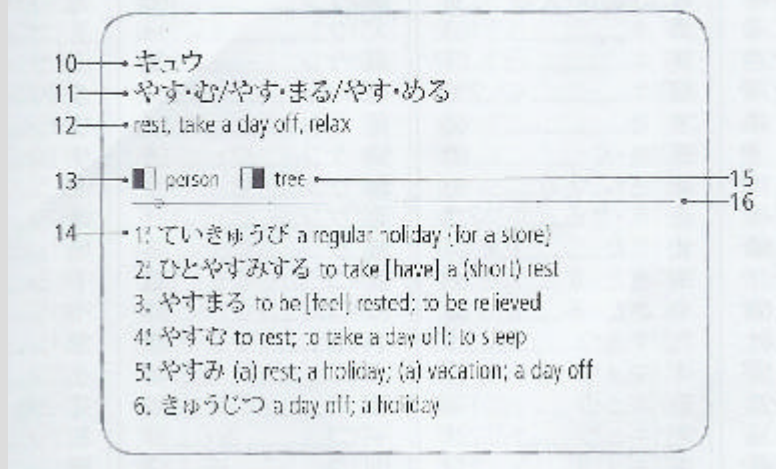
225/55/16 V

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)

Learning object: Kanji Flashcard ?? ?

Front side: kanji, examples and additional information →



← Back side: lexems, pronunciation, meanings and additional information

Source: Whiterabbit kanji flashcard

→ CLIL – Content and Language Integrated Learning

Entities of structured content

■ Simple entities of structured content

- Lexicographical entry
- Terminological entry
- Etc.

■ Complex entities of structured content

- Addresses in a database
- Learning object at the level of lexical semantics
- Extended nomenclature entry
- Product catalogue / property / classification entry
- Patient record
- Animated sign language cartoons
- Etc.

→ content products and services = business

STRUCTURED CONTENT TODAY

■ Terminology

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

■ Thesauri, classification schemes, keywords

■ Encyclopedic (knowledge) entries

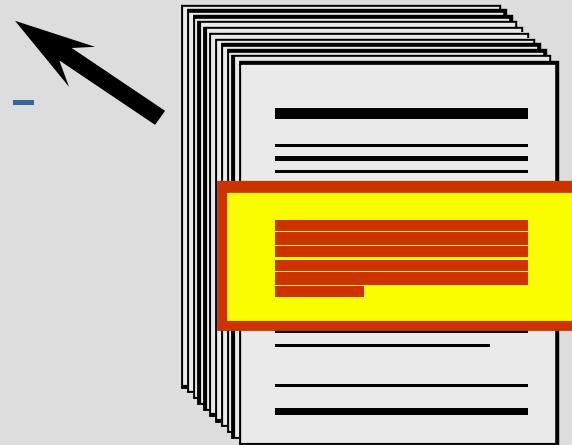
- Knowledge-enriched terminology entries
- (explained) proper names, ...

■ Ontologies, topic maps, ...

→ often contradictory, NOT coherent, integrated, reliable, ...

Updating Item Descriptions

“Fastened by a
steel 3-1/2”
threaded bolt”



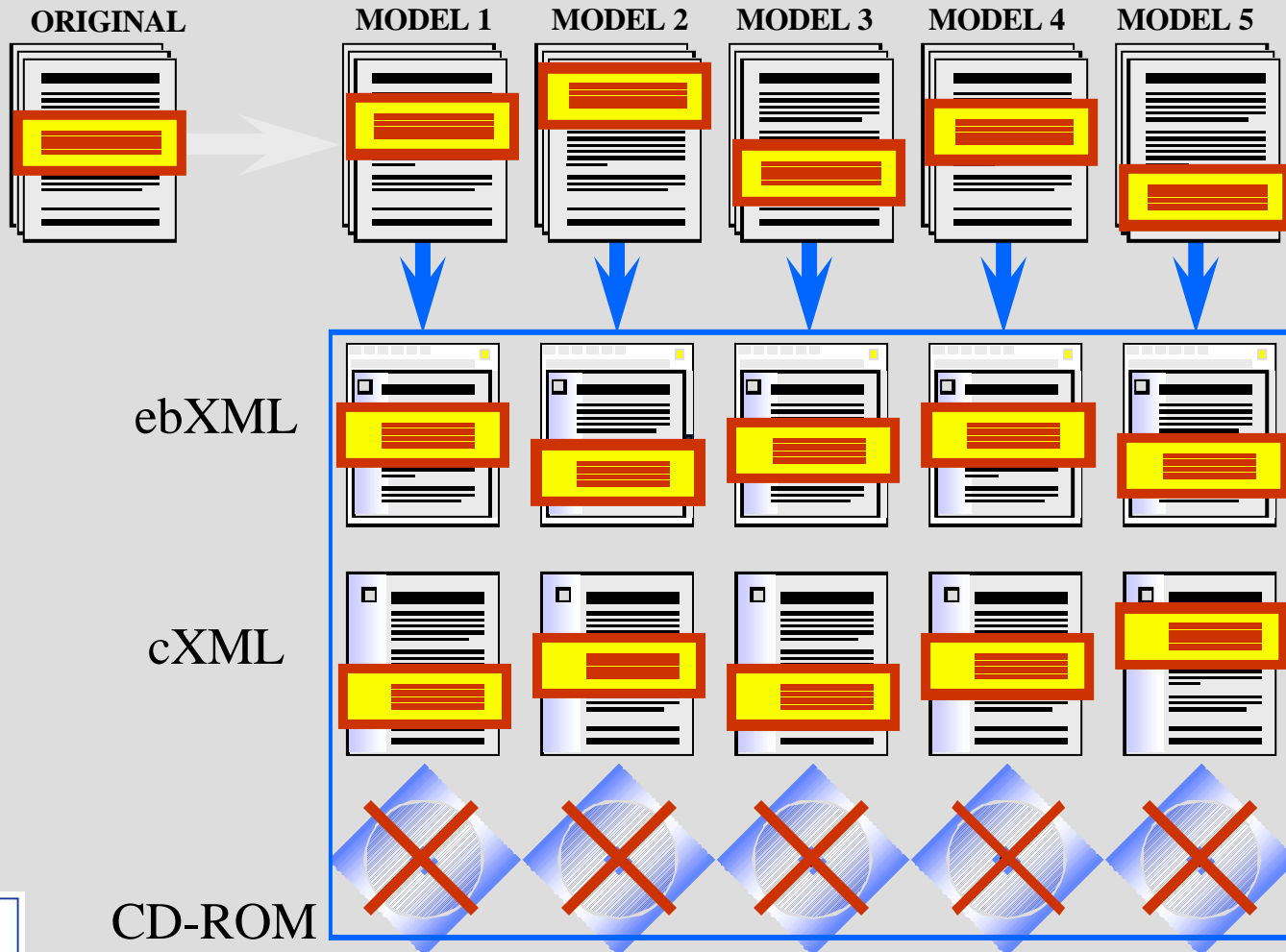
“Fastened by an
aluminum 3-1/2”
threaded bolt”

Source: Ben Martin (J.D. Edwards) 2002

Updating Item Descriptions

Out to different exchanges and formats

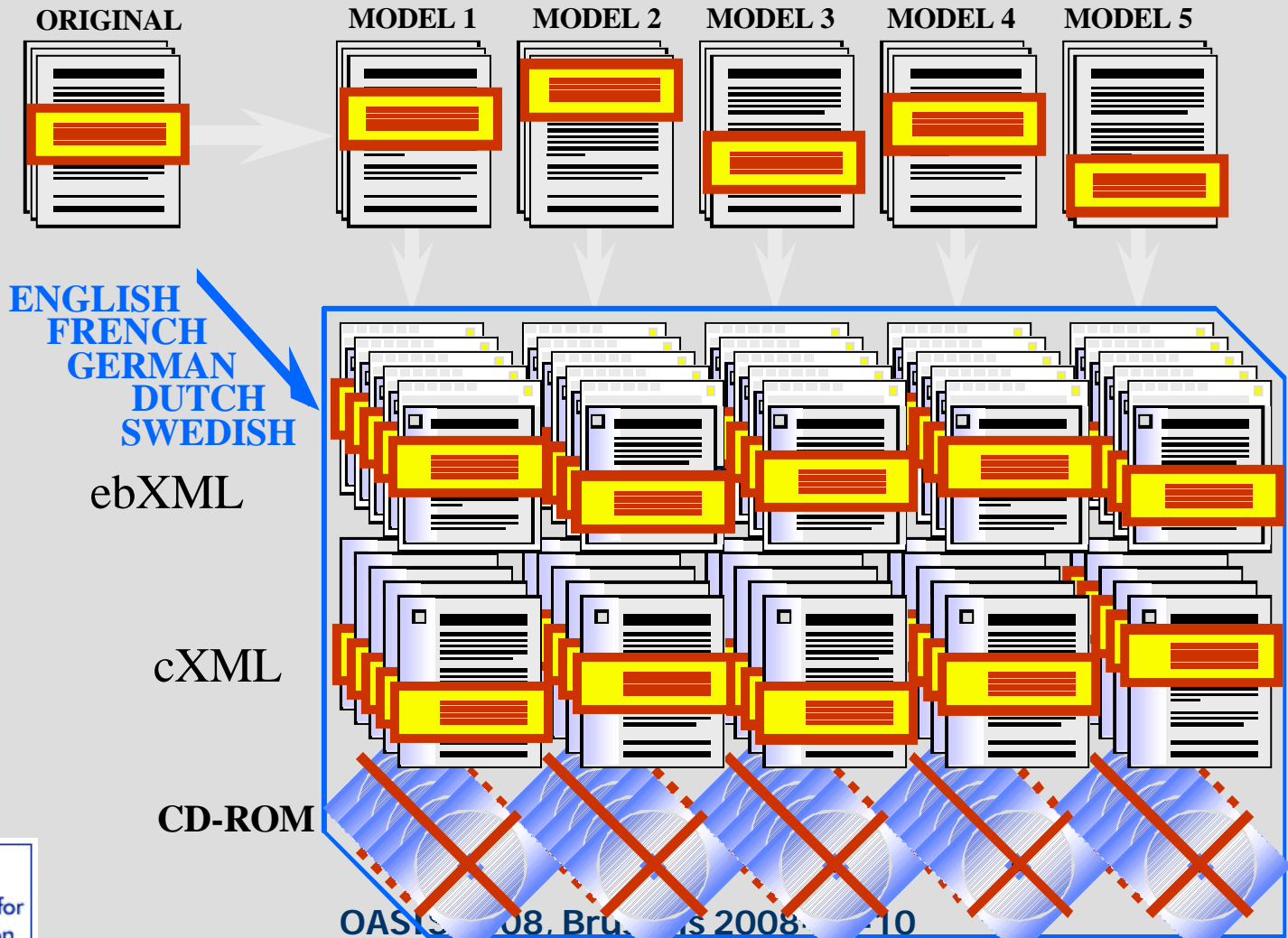
Source: Ben Martin (J.D. Edwards) 2002



Updating Item Descriptions

... ideally into various languages

Source: Ben Martin (J.D. Edwards) 2002



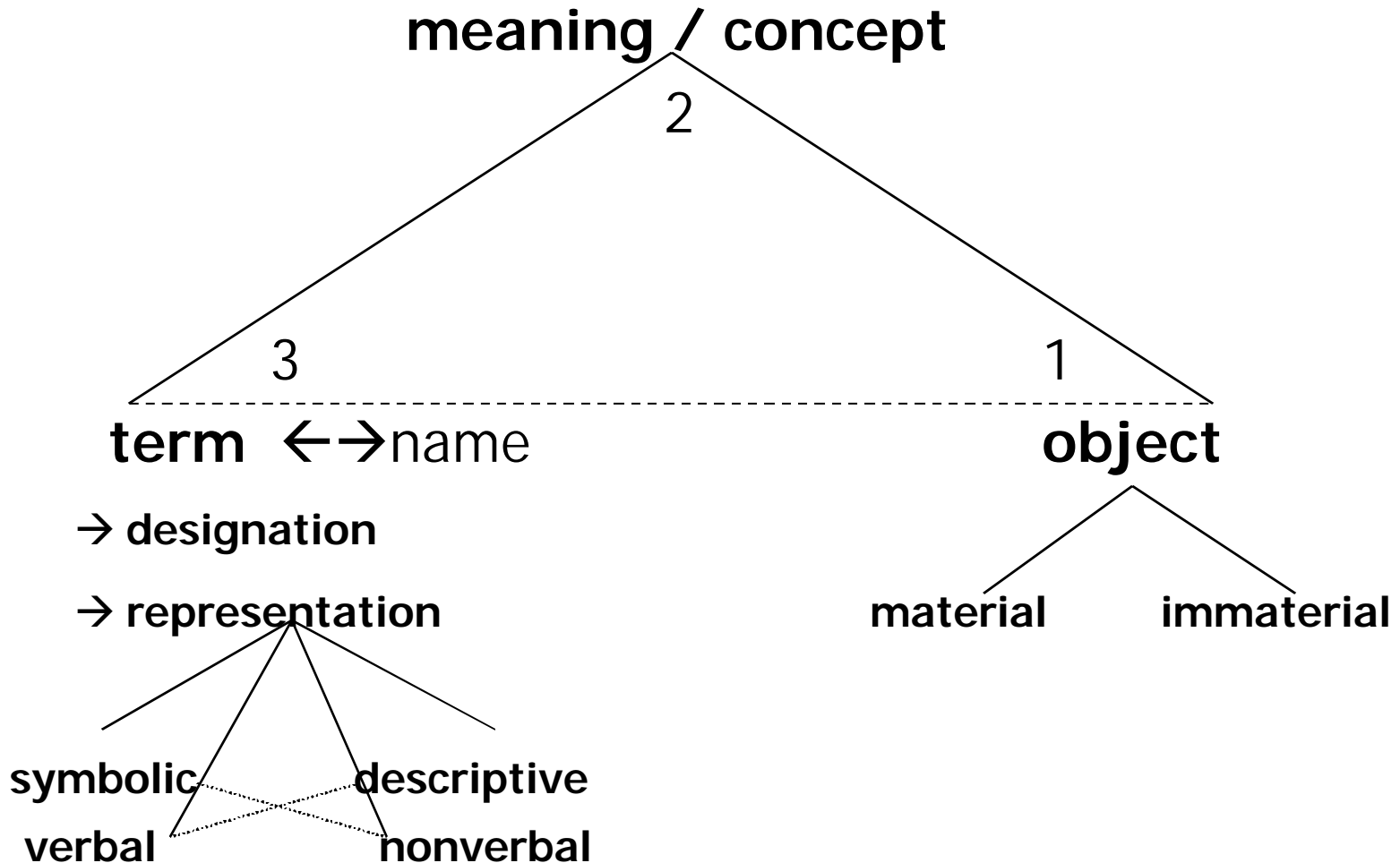
Terminology \leftrightarrow e/mContent

- **embedded terminology** (or combination of terminology + ...)
 - Terminologies as means of domain-specific communication
 - Terminology as knowledge representation
 - Terminology indispensable for domain-specific education and training
 - Terminologies as means of access to other kinds of information (objects)
 - Terminologies as means of knowledge ordering at micro-level
- **Multifunctional nature of terminology**
- **Knowledge-rich terminology**
 - Encyclopedic knowledge: Wikipedia...—
 - “Knowledge” management: \rightarrow incl. true “content management”
 - document management,
 - communication management,
 - information management
- **“popularized” terminology**

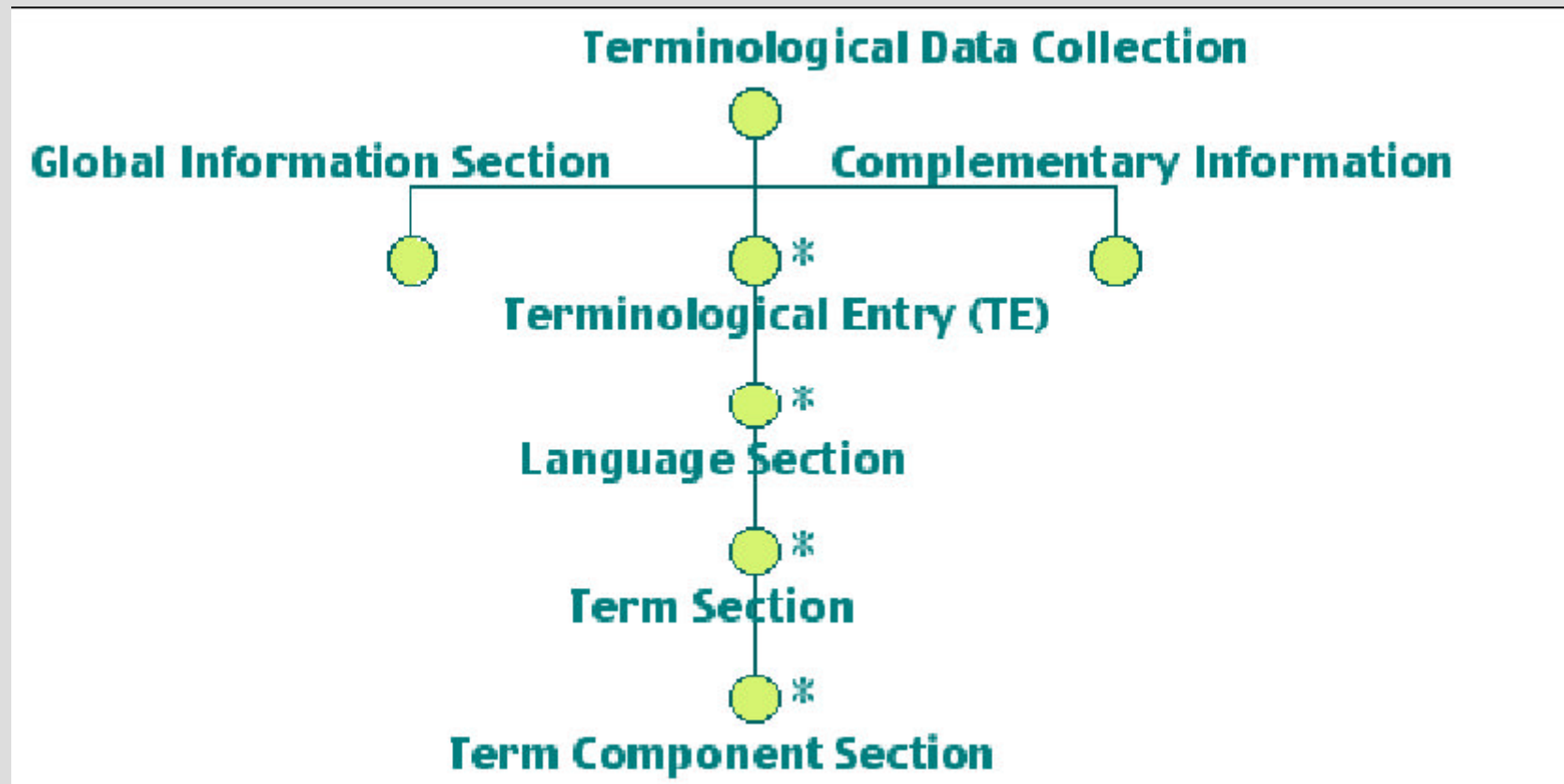
+non-verbal representations

- \rightarrow ISO/TC 37 “Terminology and other language and content resources” provides a coherent combination of methods**
- \rightarrow ONE methodology**

The "semantic triangle"



Terminological metamodel



Source: Klaus-Dirk Schmitz 2005

+ OTHER CONTENT RESOURCES

■ Non-verbal representations

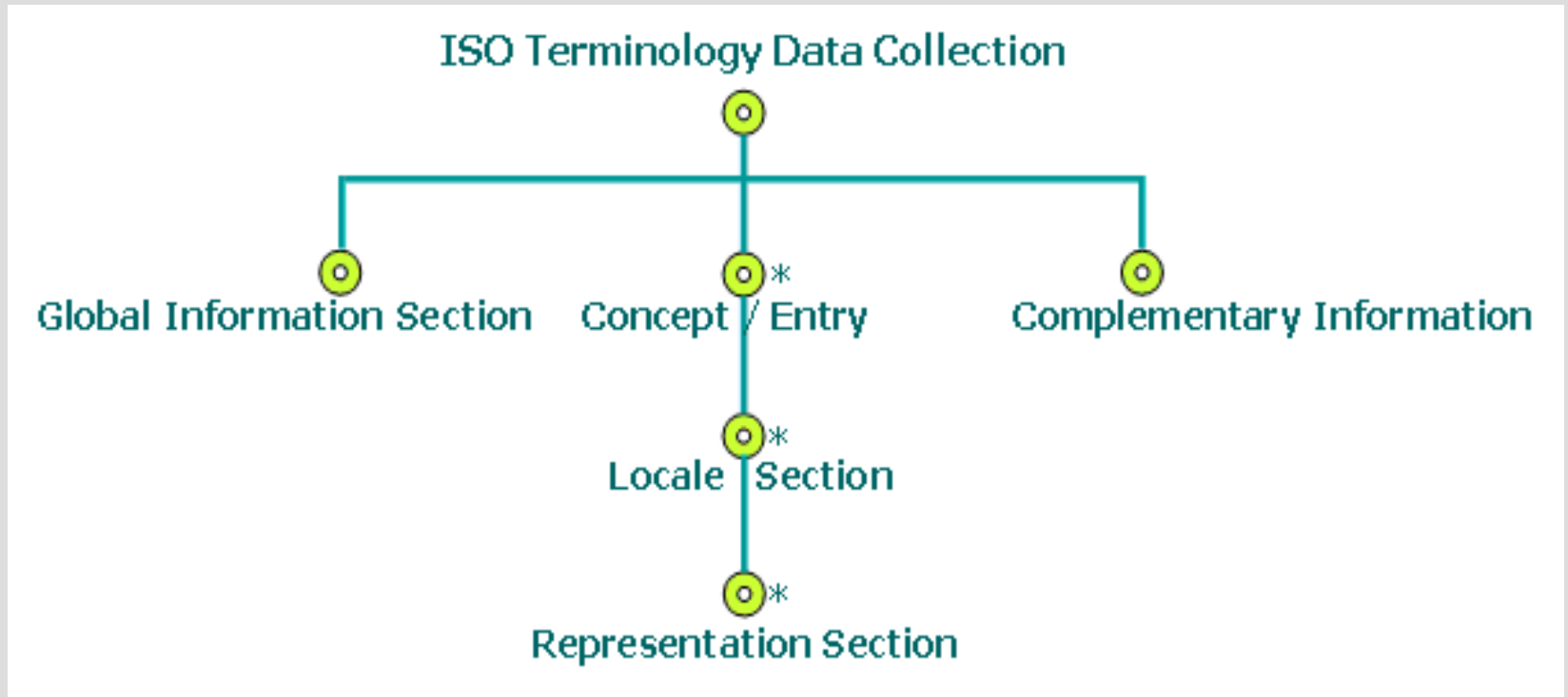
- Visual: symbols, graphics, ...
- Haptic
- Etc.

■ Non-verbal communication

- Mimics
- Gestures
- Etc.

→ **locales!**

Terminological Metamodel revised



Source: Klaus-Dirk Schmitz 2005

Increasing availability on the web

BUT...

- Problem 1: **quality of content** – reliability, authoritativeness, lack of certain data, ...
- Time and place: every information any time everywhere, but ... –
- Problem 2: **redundancy/duplication**
- Problem 3: **inconsistent and incoherent**
- Problem 4: **not fit for purpose**
- Etc.

Nevertheless access as such is less and less a problem

Distribution via the web

- Accessibility in the web = access
 - **Distribution via the web: easy, BUT**
 - **Extensive re-use** of existing data
 - more garbage
 - more redundance
 - copyright issues
 - **Extensive R&D in making intelligent re-use**
 - **New distribution methods and channels**
 - **But: lack of sustainable new business models**
- nevertheless, the problem is decreasing from technical p-o-v
- Creation through the web

THE problem: DEVELOPMENT - CREATION AND MAINTENANCE

- Accessibility in the web = access
- Distribution via the web
- **Creation and maintenance through the web**
 - Semantic web? → social web
 - Cooperative & distributed approaches
 - new kinds of ownership
 - New workflow methods **+ standards**
 - Federation of repositories **+ standards**
 - Needs constant and continued efforts
 - Many new aspects **+ standards**

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

FEDERATED CONTENT

- **Repositories of different types of content**
→ based on metadata (& metadata registries)
- **Content repositories for defined domains/applications**
→ maintained by the community best fit for
- **Different language versions of such repositories**
→ maintained by the language community interested
- **Federating different types of content**
→ semantic interoperability → content interoperability

federated repositories ↔ content interoperability
→ content interoperability standards!

IMPACT

- Terminology science and its applications are under pressure for change
- The same applies to
 - Applied linguistics
 - Computational linguistics
 - Communication science
 - Content management
 - Knowledge management
 - →even science theory
- alignment of several sciences/theories and their applications →content interoperability
→content interoperability standards

Thank you for your attention

ISO/TC 37 Secretariat

*(on behalf of the Austrian Standards
Institute – ON)*

ADDRESS

CHAIRMAN: Håvard Hjulstad (SN)

SECRETARY: Christian Galinski

ISO/TC 37

**c/o Infoterm – International Information
Centre for Terminology**

Mariahilfer Strasse 123/3

A-1060 Vienna – Austria

Tel: +43-664-344 6181

Fax: +43-1-5876990

infopoint@infoterm.org

<http://www.infoterm.info>

TERMINOLOGY STANDARDIZATION

- Hitherto prominently verbal-linguistic term-oriented approach to terminological data management needs to take non-verbal 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

Distributed cooperative terminology standardization

Proposal: Wiki-type work group system platform (supplemented by additional information system components):

- *discussion forum*
- *easy-to-access bibliographic database*
- *moderation group*
- *editorial group*
- **terminology data bank** (such as the planned ISO/CDB) with the following characteristics:
 - history to many/most of the text fields
 - extensive source information
 - LOG file reflecting the reasoning behind decisions in the process of terminology standardisation
(to be kept by the respective committee in order to save most of the time wasted on recurring discussions in later stages of terminology standardisation)
- **administration database with factual data on experts (their addresses and contact details, etc.), related institutions / organisations / projects, etc.**
- **other tools for the secretary/secretariat of the respective project.**

Discussion forum

- open to all (but contingent upon prior registration and in accordance with certain rules of good conduct)
- to discuss a limited number of closely related terminology entries (possibly in one lead language – not necessarily English)
- on the basis of terminological principles and methods
- system-supported from the very beginning of the project (without making things too difficult for the experts);

Bibliographic database

- *easy-to-access BDB*
- with information on “qualified” sources suggested by the members of the discussion forum
- via a quality evaluation and verification/validation process supported by the system
- bibliographic entries to be made by anyone according to established rules,
(computer-assisted, which would constitute a “validation and qualifying process” for suggested sources)
- **Coded sources + URI system**

Moderation Group

- composed of a number of selected field experts (partly drawn from the respective TC and also comprising some of the editors)
- to analyse and summarise (and evaluate) the essence of the contributions in preparation for decision making
- Comp.: ISO 639/RAs-JAC

Editorial Group

- made up of decision makers from the respective TC
- to introduce consolidated entries (equivalent to CD stage) on the basis of defined
 - ToRs (terms of reference)
 - RoPs (rules of procedure) -into the terminology collection of the TC
- in preparation of a final decision on which entries will be integrated into the respective IS or ST collection
(other entries could/should be kept as “non-standardised” for future reference)

+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 & metamodel(s)**
- **Exchange / interoperability**
- **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)**: definition, explanation, non-linguistic [descriptive] representation, 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 content entities**

Semantic 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...)
- ...

STANDARDS for METHODOLOGY & DATA

Top-down vs. bottom-up approaches

Standardization – Top-down → the rules

- Harmonization of metadata
- Unification of principles and methods of data modelling
- Standardization of meta-models
- Standardization of workflow methodology

Standardization – Bottom-up → the data

- Product classification
- Product identification
- eCatalogue data
- ontologies
- terminologies
- LRs

by using net-based distributed cooperative working methods – peer2peer

METHODOLOGY

APPLICATIONS

<p>ISO 16642*</p>	<p>(family of) metamodels*</p>				
<p>Datamodels ISO 12200**</p>	<p>Datamodels** eBusiness</p>	<p>Datamodels other e...s**</p>	<p>Datamodels other e...s**</p>		
<p>Data categories ISO 12620***</p>	<p>Domain data dictionaries***</p>	<p>DDD ***</p>	<p>DDD ***</p>	<p>DDD ***</p>	<p>DDD ***</p>

Basic principles and requirements concerning multilingual e/m-content development, data categories/metadata, data modelling, rules for repositories (maintained in MAs/RAs/Reg's)

ISO 16642 TMF; ISO 10303-11 EXPRESS; ISO 10303-21 SDAI; ...

ISO 12200 MARTIF; ISO 13584-42 PLIB ~ IEC 61360-2

*ISO 12620 Data categories; ISO 13584-511 Fastener dictionary; IEC 61360-4 Core dictionary; ...

ISO/IEC policy concerning repositories

Required is a systematic approach to

- maintenance agencies
- registration authorities
- other registries
- agreements for re-use (federated business model also incl. solutions for copyright issues)
- coordination of pertinent standardization efforts

→ *ISO & IEC to develop a comprehensive policy and a coherent strategic framework of federated repositories for the sake of global semantic interoperability*

WHAT IS TERMINOLOGY?

- The description of the specialized vocabulary of an application domain
 - Terminology takes the conceptual view
→ knowledge representation at concept level
 - Monolingual or bilingual? → multilingual
 - Mainly nouns (incl. multi-word nominal units)?
some verbs, adjectives and adverbs
 - Terminology: a strong yet practical simplification of lexical description (L. Romary)

Increasingly different (technical) types of content co-occur or are embedded in each other or are combined with each other – e.g. in "traffic telematics"

Terminology \leftrightarrow Web

- **Accessibility in the web = access**
 - by anybody from everywhere
 - in the form and at the time needed
 - fit for the purpose required
 - **incl. people with special needs** (universal ~)
- Distribution via the web
- Creation through the web

Driving force: ICT development

- Internet → Web 2.0
→ social web
- Web 3.0 (or later?)
→ semantic web
- Web 4.0
→ mobile web
- Web 5.0
→ ubiquity and pervasiveness

eContent

= digital content

→ mContent

= mobile content

or does it happen simultaneously?

Software development in general is still poor if requirements of content integration and interoperability are considered!

G11N – L10N – I18N

Closely related and highly interdependent concepts:

- Globalization – G11N
→ lots of misconceptions
- Localization – L10N –
nothing more than ‘high-tech translation’?
- Internationalization – I18N
→ makes L10N efficient and cost-effective!

→ **GLOCALIZATION**

LOCALIZATION

- **LISA 2007:**
 “Localization is the process of modifying products or services to account for differences in distinct markets” [...] localization must be integrated with other business processes if it is to be effective. Localization is an integral part of globalization, and without it, other globalization efforts are ineffective.”
 - **EURESCOM 2000:**
 Key concept: *locales*, which refers to a collection of people who share language, writing system and any other properties which would require a separate version of a product .”
- replace product by structured content and related services

GLOBALIZATION

- **Kofi Annan:**
G11N is “an irreversible process, not an option”
- **LISA 2007:**
“the only way for enterprises to be global is to be simultaneously local in the markets in which they do business. By respecting local languages and culture at every level – in their products, services, documentation, customer support, marketing, maintenance procedures, business practices, etc. – global enterprises paradoxically expand the options available at the local level.”

INTERNATIONALIZATION

LISA 2007: *Internationalization*

- ***“...is the process of enabling a product at a technical level for localization”*** → including – of course – also services
- enables a product or service not to require remedial engineering or redesign at the time of localization;
“Localization does not just happen; it must be planned for. Even in the simplest cases, there will be issues that will work against successful localization:
 - graphics may contain embedded text that must be translated
 - screenshots may appear in a particular language
 - colours may be culturally significant
 - phone numbers may be usable only in one country
- **Because of such issues, a process known as I18N is used to remove cultural assumptions from products during development so that they can effectively be localized.”**

TERMINOLOGY: STRATEGIC ASPECTS

Given its pervasive occurrence in all (written or spoken) domain communication, terminology today has to be considered an economic factor especially in

- product data description and management (incl. eCatalogues and product classification)
- quality management
- inter-cultural aspects of management and marketing
- translation and localization
- information, documentation, software development
- knowledge transfer...

→ ***terminology science***

- *is a field of fundamental research as well as applied R&D*
- *applies to other kinds of structured content*

→ ***impact on standardization***

ISO/TC 37 – SCOPE EXTENSION

Step 1: + 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
- **As is the case with terminologies, language resources in general have to be considered as multilingual, multimedia and multimodal from the outset.**

Scenario

You are a beginner in a given application e.g. as

- an apprentice, trainee, etc.
- job newcomer: new hire, job transfer, new profession, etc.
- domain student, researcher, etc.

whether in education, research, business, publication or any other field of new professional engagement.

What are your problems?

- understanding new terms or terms with a new/different meaning, abbreviations, etc.
- positioning new facts in some more or less unfamiliar domain knowledge structure,
- memorizing lots of new names,
- memorizing lots of new facts,
- understanding the new community “culture”, etc.

Access to terminology

Wouldn't it be nice, if things would be explained

- In communication with unambiguous and consistent terminology?
- In textbooks with consistent and well defined terminology?
- In works/tools of reference with reliable information?
- On self-learning tools (e.g. learning flashcards) with coherent cross-references?
- Not only in a foreign language, but also in your language?
- Taking care of your learning preferences, handicaps, etc.?

Terminology + names + graphic symbols + facts + ...

CONSEQUENCES

- Terminology science and its applications are under pressure for change
 - The same applies to
 - Applied linguistics
 - Computational linguistics
 - Communication science
 - Content management
 - Knowledge management
 - → even science theory
- alignment of several sciences/theories and their applications → content interoperability

Terminology & content integration I

- 1950s: terminology standardization without proper theoretical and methodological foundation; dominance of domain experts in terminology
- 1960s: huge undertakings in I&D (information and documentation); early beginnings of terminology theory
- 1970s: dominance of I&D experts in terminology standardization; early beginnings of LSP; development of theoretical and methodological foundation of terminology standardization;

1971 Infoterm founded by UNESCO

Terminology & content integration II

- 1980s: dominance of LSP experts and translators in terminology standardization;
 - emergence of terminology science;
 - large T&D departments;
 - explicit language policies;
 - „big“ translation-oriented TDBs;
 - stabilization of the theoretical and methodological foundation of terminology standardization;
 - big efforts in machine translation;
 - discussion on „professionalization“ of translation, technical communication, etc.;