ISO/TC 37

Terminology and other language resources management standards and application

CHOI, Key-Sun
NHK STRL
KAIST Korterm
(Korea Advanced Institute of Science & Technology)
(Korea Terminology Research Center for Language & Knowledge Engineering)
ISO/TC 37/SC 4 Secretary
http://korterm.org/

TC 37 history

- ISO/TC 37 N410 (Rev.) and TC 37/AG N26
- 1936 founded as ISA/TC 37 “Terminology” in the International Federation of the National Standardizing Associations (ISA)
  - Worked until 1939
- 1951 ISO/TC 37 “Terminology (principles and co-ordination” of the International Organization for Standardization (ISO)
  - Started to operate in 1952.

SC

- SC1 Principles and methods
  - Principles of terminology (until 2001)
- SC2 Terminography and Lexicography
  - Layout of vocabularies
- SC3 Computer applications for terminology
- SC4 Language resource management
  - (start from June/2002, Las Palmas)

Terminology Standards

- Two meanings of “Terminology Standard”
  - “Vocabularies”
    - Terminology standards that contain subject-field-specific concepts and terms produced by terminology sub-committees on national, regional and international level
  - Terminology-principles-and-methods standards
    - Produced by specific committees on national and international level (ISO/TC37, JISC-NAT, …)

Vocabularies (Principles and Methods)

ISO 1087-1

Terminology – Vocabulary – Part 1

ISO 1087-2

Terminology work – vocabulary – Part 2: Computer Applications

ISO 1087-1 contents

Contents

Foreword iv
Introduction v
1 Scope 1
2 Normative references 1
3 Vocabulary 2
3.1 Language and reality 2
3.2 Concepts 2
3.3. Definitions 5
3.4 Designations 5
3.5 Terminology 8
3.6 Aspects of terminology work 9
3.7 Terminological products 10
3.8 Terminological data 11
Annex A (informative) Concept diagrams 13
Annex B (informative) Alphabetical index 15

ISO 1087-1 “Vocabulary”

3 Vocabulary

3.1 Language and reality

3.1.1
object
anything perceivable or conceivable

NOTE Objects may be material (e.g. an engine, a sheet of paper, a diamond), immaterial (e.g. conversion ratio, a project plan) or imagined (e.g. a unicorn).

3.1.2
subject field
domain
field of special knowledge

NOTE The borderlines of a subject field are defined from a purpose-related point of view.

3.1.3
special language
language for special purposes
LSP
language used in a **subject field** (3.1.2) and characterized by the use of specific linguistic means of expression

NOTE The specific linguistic means of expression always include subject-specific **terminology** (3.5.1) and phraseology and also may cover stylistic or syntactic features.

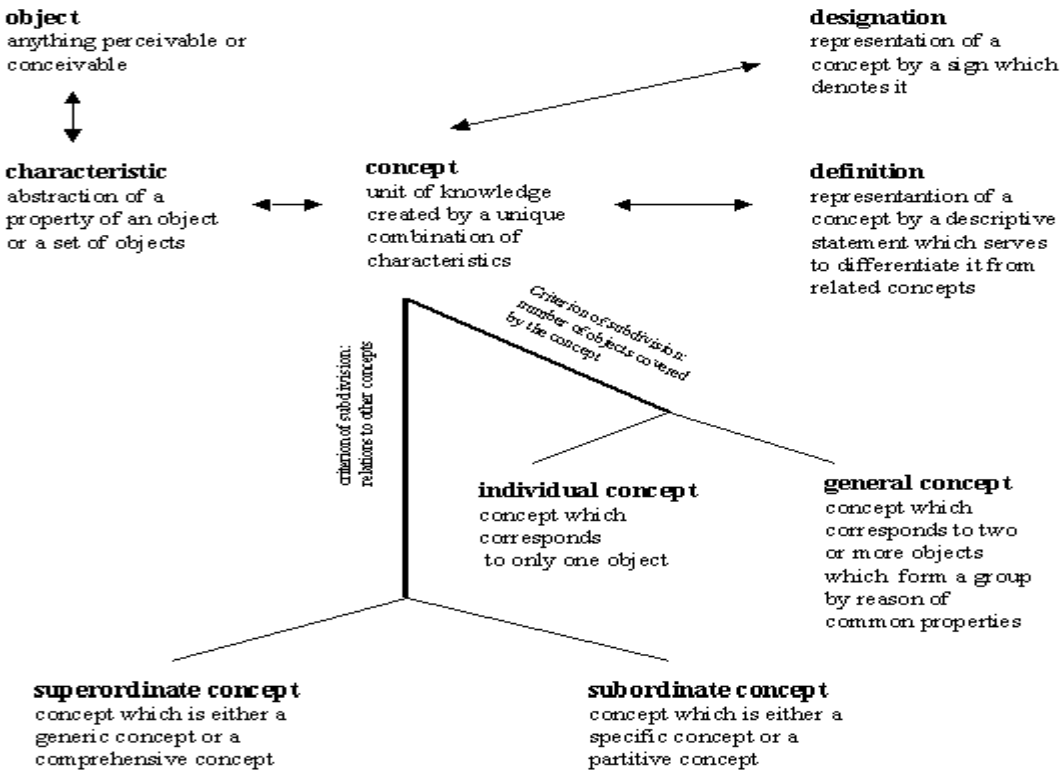
3.2 Concepts

3.2.1
concept
unit of knowledge created by a unique combination of **characteristics** (3.2.4)

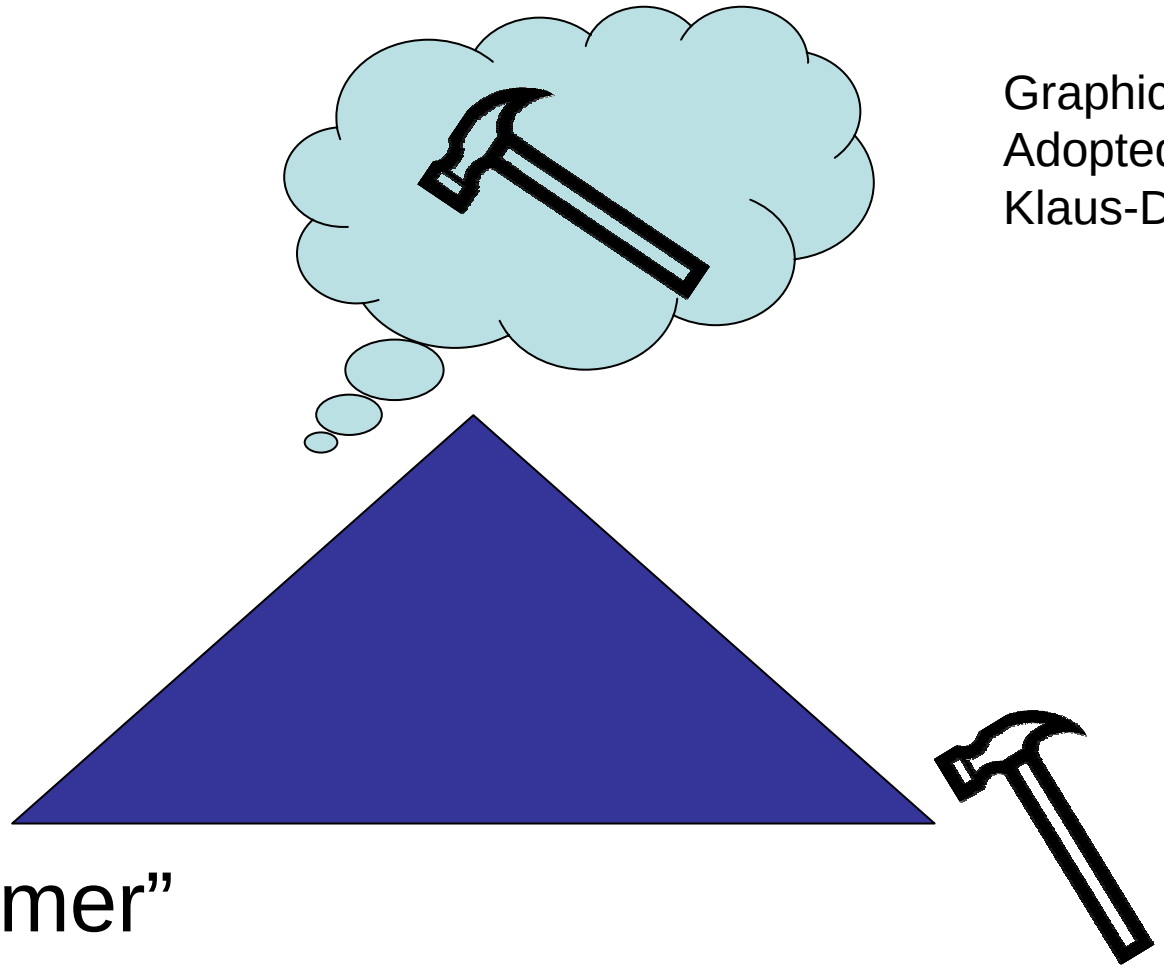
NOTE Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background often leading to different categorizations.

ISO 1087-1 Language and Reality

A.2 Language and reality



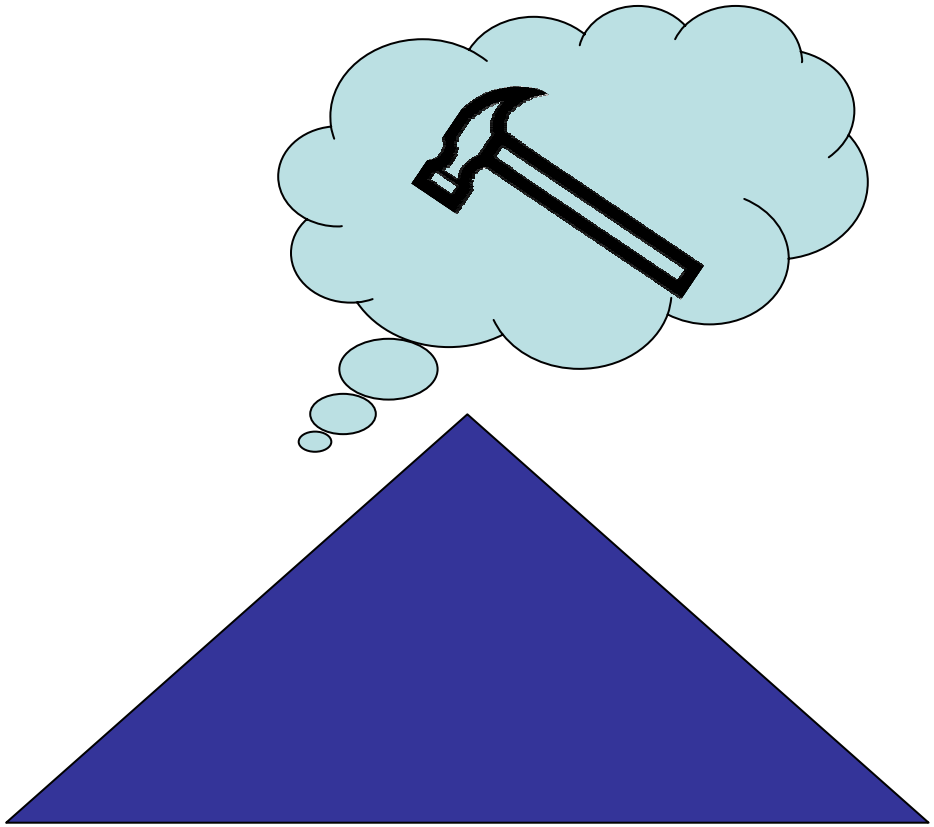
Concept Orientation: Language and Reality



Graphic series
Adopted from
Klaus-Dirk Schmitz

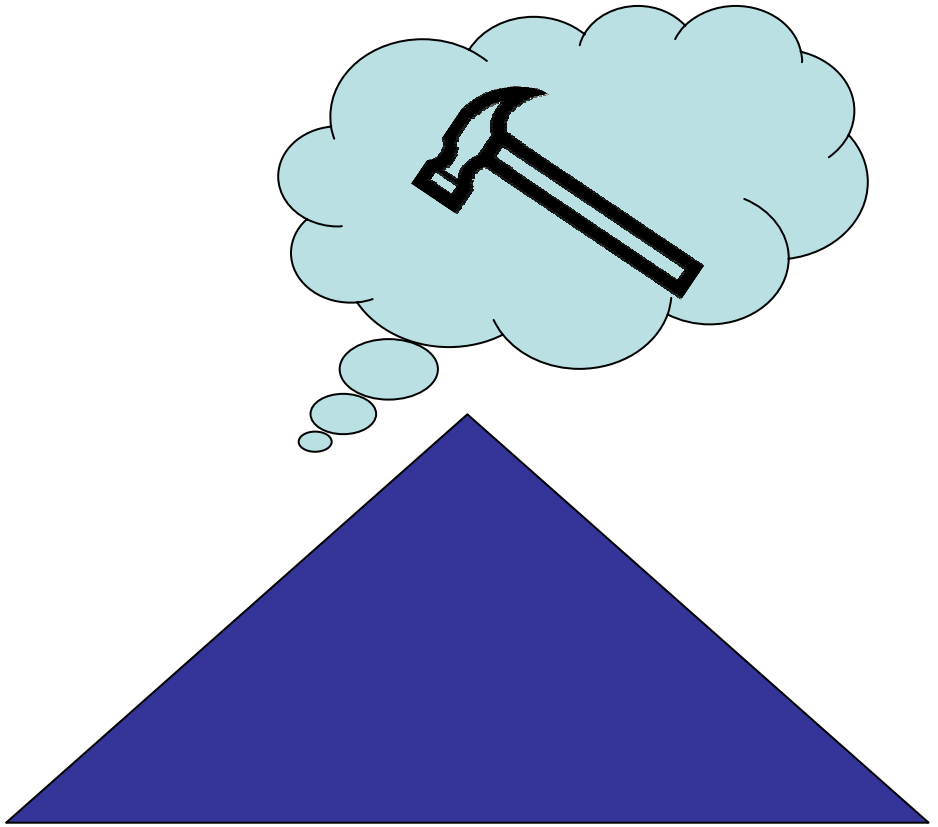
“hammer”

Concept orientation: object



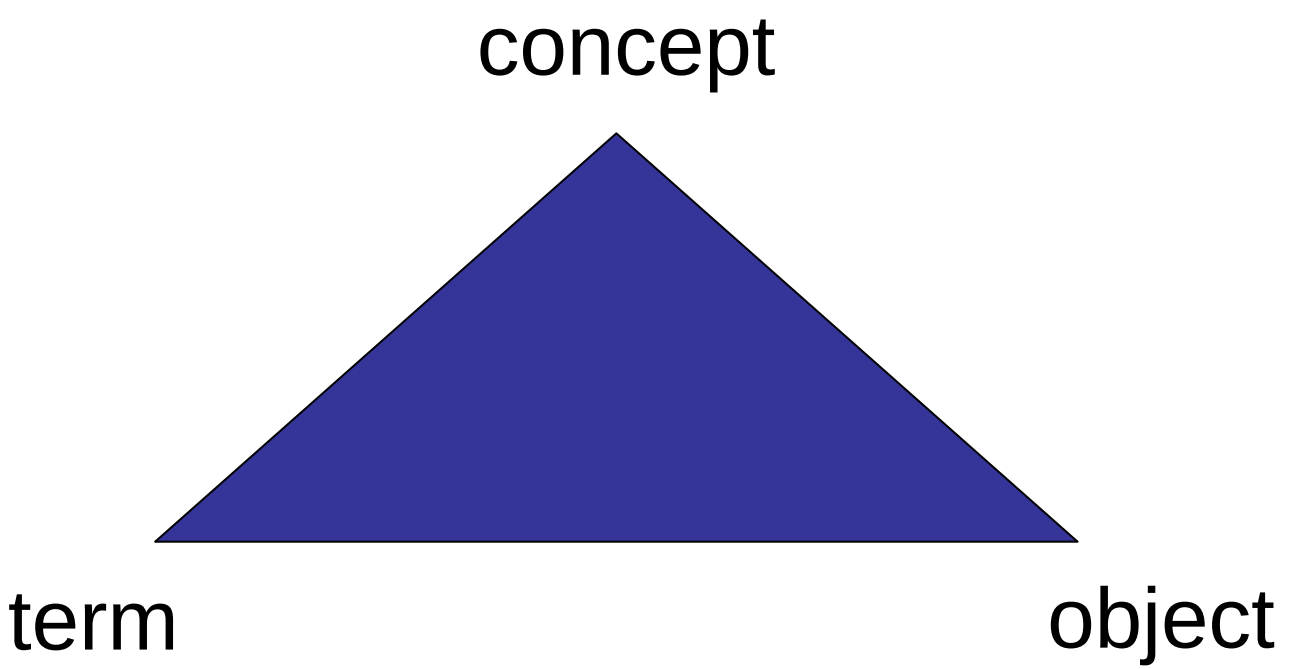
“hammer” object

Concept orientation: term



term object

Concept orientation: concept

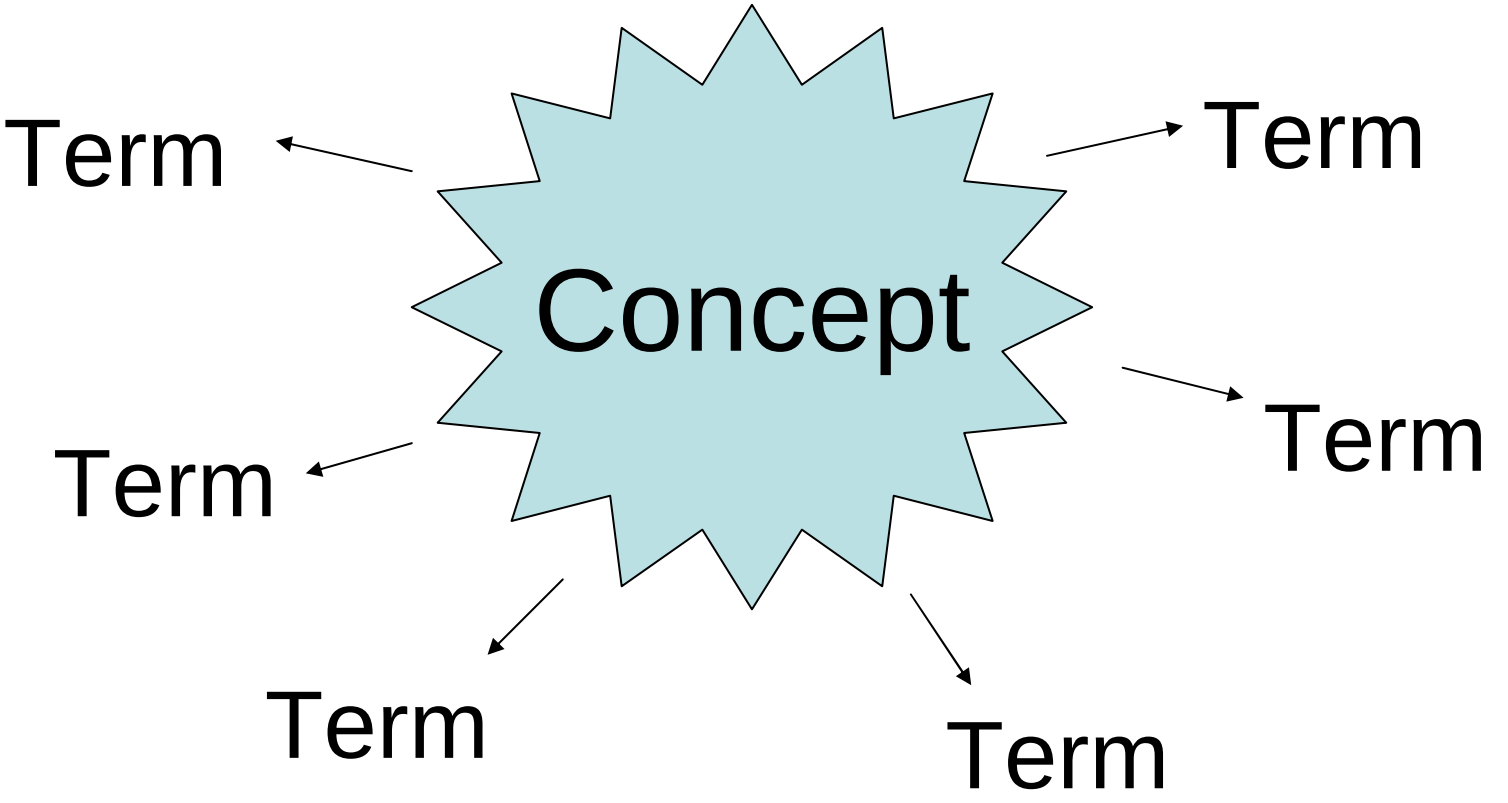


Concept orientation: definition

- Object
  - Any part of the perceivable or conceivable world
    - Objects may be material (e.g., engine) or immaterial (e.g., magnetism)
- Concept
  - Unit of thought made up of characteristics that are derived by categorizing objects having a number of identical properties
    - Concepts are not bound to particular languages. They are, however, influence by social or cultural background.
- Term
  - Designation of a defined concept in a special language by a linguistic expression
    - A term may consist of one or more words.

Terminological entry

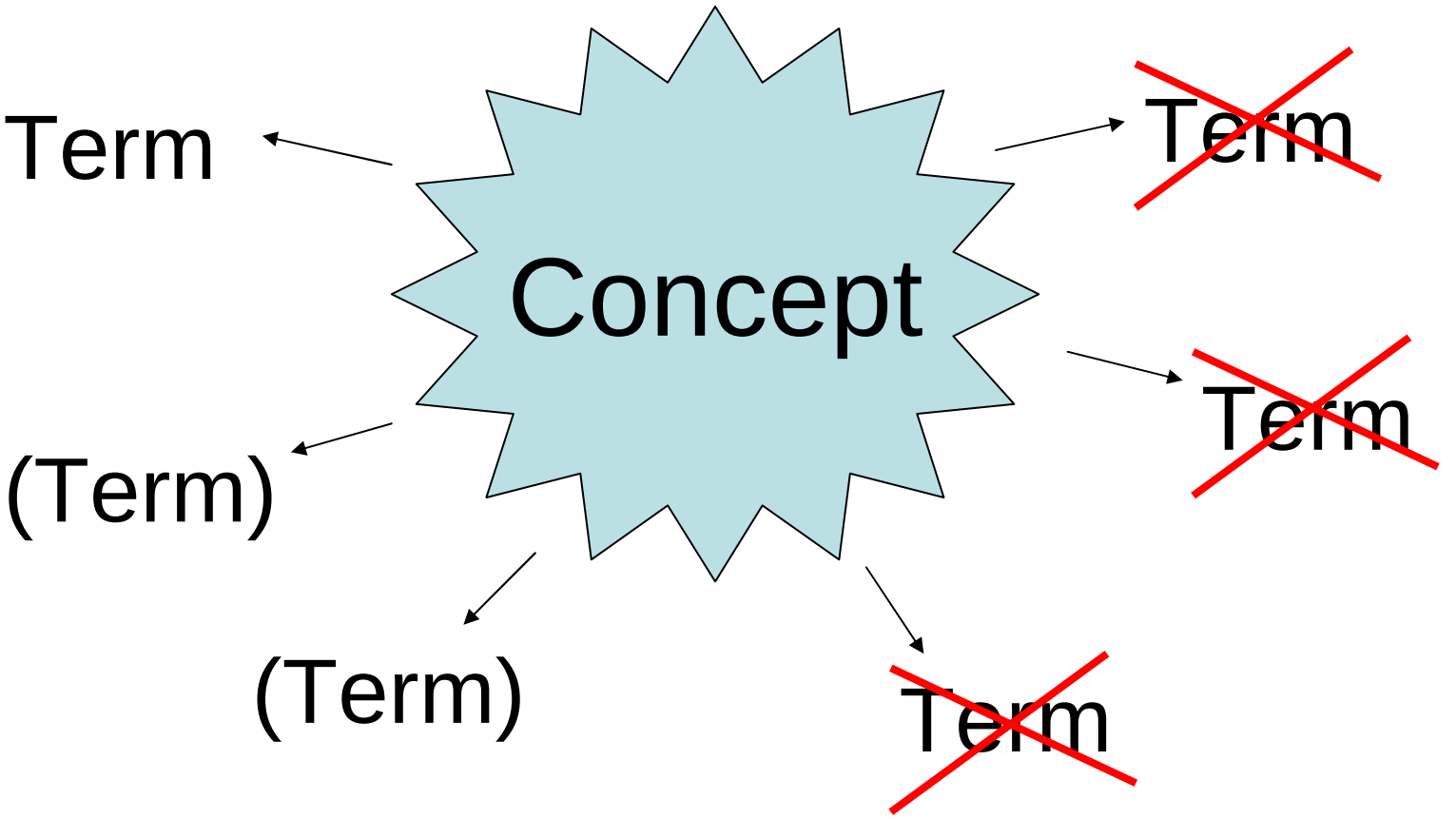
Graphic adopted from
Sue Ellen Wright and
Klaus-Dirk Schmitz



Terminological Entry: standardized

In standardized terminology:
only one (preferred) term !

Graphic adopted from
Sue Ellen Wright and
Klaus-Dirk Schmitz



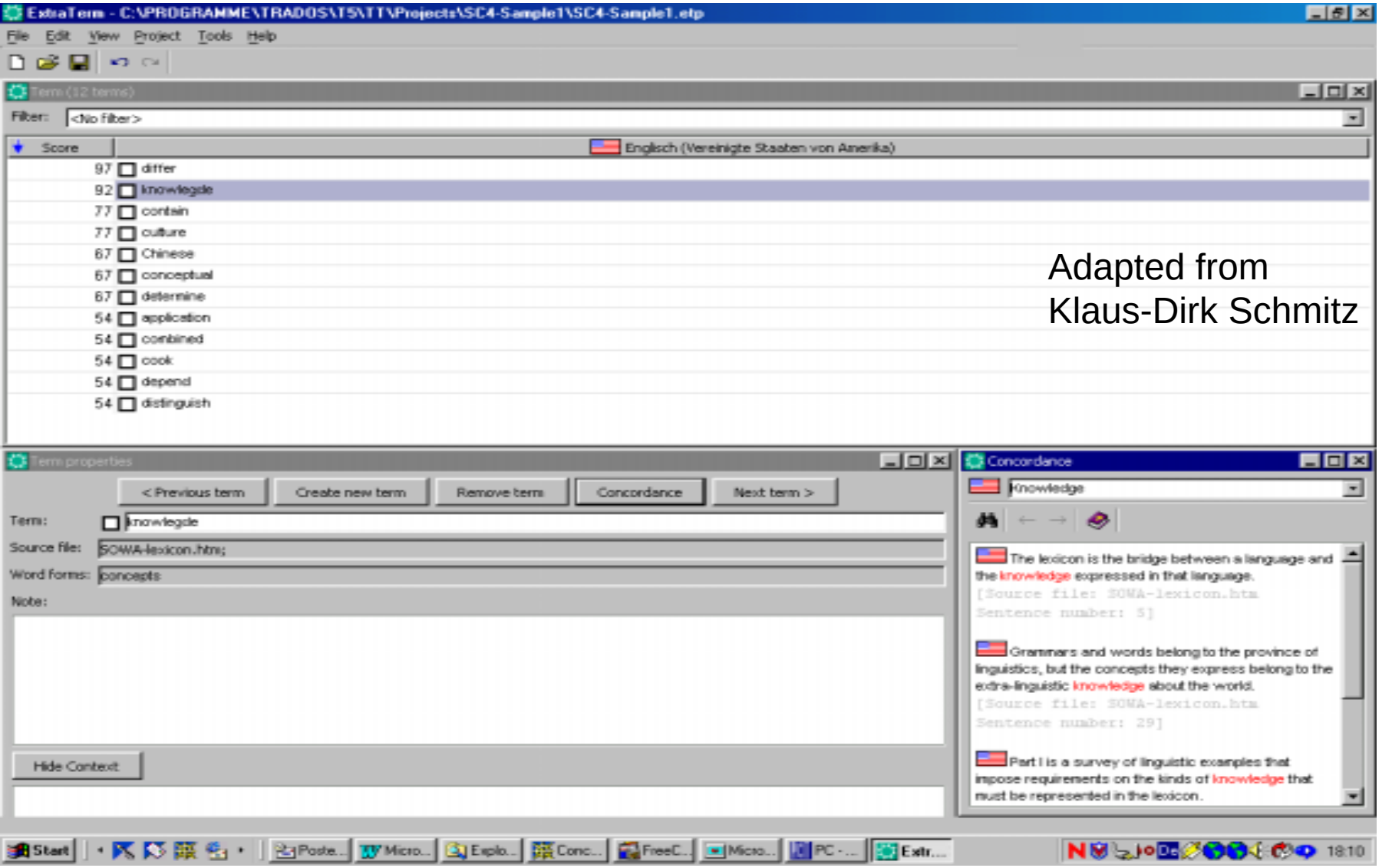
Concepts in the Lexicon: Introduction

John F. Sowa

http://users.bestweb.net/~sowa/ontology/lexicon.htm

- The lexicon is the bridge between a language and the knowledge expressed in that language. Every language has a different vocabulary, but every language provides the grammatical mechanisms for combining its stock of words to express an open-ended range of concepts. Different languages, however, differ in the grammar, the words, and the concepts they express. The differences arise from three kinds of variation:
- *Accidental*. The most obvious differences result from arbitrary choices of sounds, such as *hand* in English and *mano* in Italian. Other variations depend on arbitrary choices of where to draw boundaries. In English, *hand* refers to the part of the body from the fingertips to the wrist. But in Russian, the corresponding word *ruka* extends all the way to the elbow.
- *Systematic*. The grammar of a language determines how the conceptual structures are linearized as strings of words in a sentence. English and Chinese, for example, put the subject first, the verb in the middle, and the object at the end for an SVO word order. Irish and Biblical Hebrew are VSO languages that put the verb first. Latin and Japanese are SOV languages that put the verb at the end. The grammar also determines how the units of meaning, called *morphemes*, are combined to form words. Chinese is an extreme example of an *analytic* language in which almost all the morphemes can be used as stand-alone words. German is an *agglutinative* language, which forms compound words like *Lebensversicherungsgesellschaftsangestellter* (life insurance company employee). Old English was an agglutinative language like German, but as it evolved into modern English, it became almost as analytic as Chinese.
- *Cultural*. The concepts expressed by a language are determined by the environment, activities, and culture of the people who speak the language. Since French, Chinese, and Indian cuisines are based on very different ingredients, methods of preparation, and cooking utensils, the people who cook and eat each kind of food use words for it that have no counterparts in the other cultures. The specialized concepts, however, can be transferred with the culture whenever a cook opens a new restaurant in a foreign land. Cultural and conceptual shifts occur across time as well as space. A book on science or business, for example, is easier to translate from modern English to modern Japanese than from modern English to the language of Shakespeare.
- Grammars and words belong to the province of linguistics, but the concepts they express belong to the extra-linguistic knowledge about the world. For each language, the lexicon must provide the links that enable a language processor to carry messages from one province to the other.

Terminology work



(Vocabularies)
Principles and Methods

ISO 12620

ISO 12200

ISO DIS 16642

ISO/TC 37/SC 3

- ISO 12620: Computer applications in Terminology – Data Categories
- ISO 12200: Computer applications in Terminology – Machine-Readable Terminology Interchange Format (MARTIF) – Negotiated Interchange
- ISO DIS 16642: Computer applications in terminology – Meta model for representing terminological data collections / Terminology mark-up framework

ISO 12620 (Data Categories)

- Inventory of more than 200 data categories used in terminological data collections:
  - A.1 term
  - A.2 term-related information
  - A.3 equivalence
  - A.4 subject field
  - A.5 concept-related description
  - A.6 concept relation
  - A.7 conceptual structures
  - A.8 note
  - A.9 documentary language
  - A.10 administrative information
  - Annex B (informative): Bibliographical data categories

ISO 12620 (Data categories) ex.

- 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

– On the basis of a study and analysis of a great variety of practical applications; can be amended

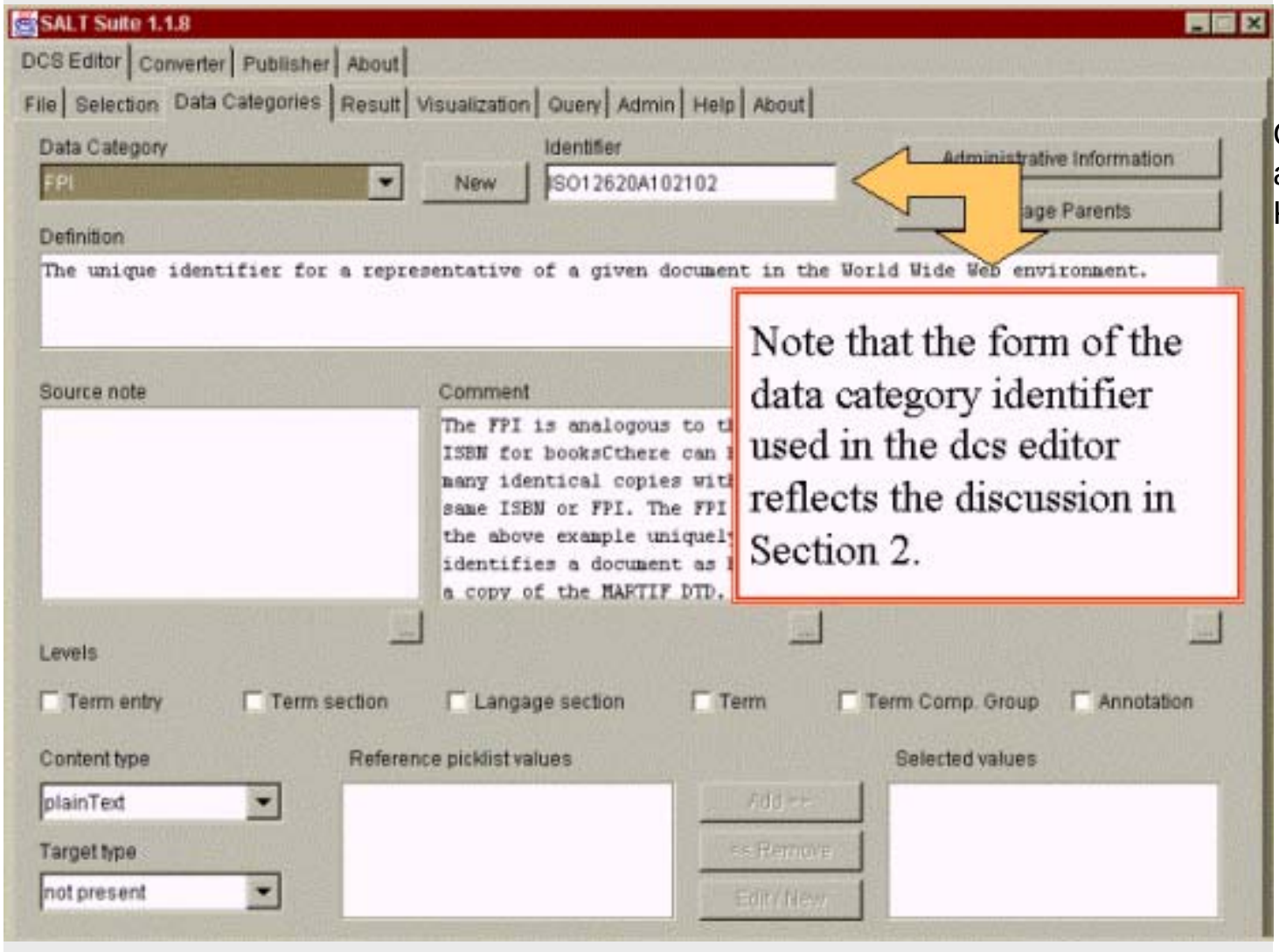
ISO 12620 new

- Metadata Registry
  - Contains terms that describe database fields
  - For describing and comparing databases
  - For human use
  - “concept-oriented” but referring to objects (fields) that are IT representations of (real) objects/concepts
  - ISO JTC1/TC32 provides a standard for metadata registries

ISO 12620 new – metadata registry

- Converting ISO 12620:1999 data category description into metadata registry format
- Using the DCS-Editor, developed within the framework of the SALT Project, for the description of the data categories
- Create the list of datCats and the description of datCats directly by the DCS-Editor as a normative annex of the new ISO 12620
- The body defines the (metadata) description format

DCS-Editor: SALT Project



Graphic adapted from Klaus-Dirk Schm

part of speech

Identifying and Definitional Attributes

Data Element ID:	ISO12620A020201 Version No: 1
Data Element Name:	part of speech
Type:	Data Element
Status:	Current 12-DEC-1999
Admitted Name:	
Non-admitted Name 1:	grammatical category
Non-admitted name 2:	word class
Definition:	A category assigned to a word based on its grammatical and semantic properties
Source-related Comment:	
Concept-related Comment:	
Example:	
Dictionary ID:	A.2.2.1

(Vocabularies) Principles and Methods - 2

- ISO 12620
- ISO 12200: Computer applications in Terminology – Machine-Readable Terminology Interchange Format (MARTIF) – Negotiated Interchange
- ISO DIS 16642: Computer applications in terminology – Meta model for representing terminological data collections / Terminology mark-up framework

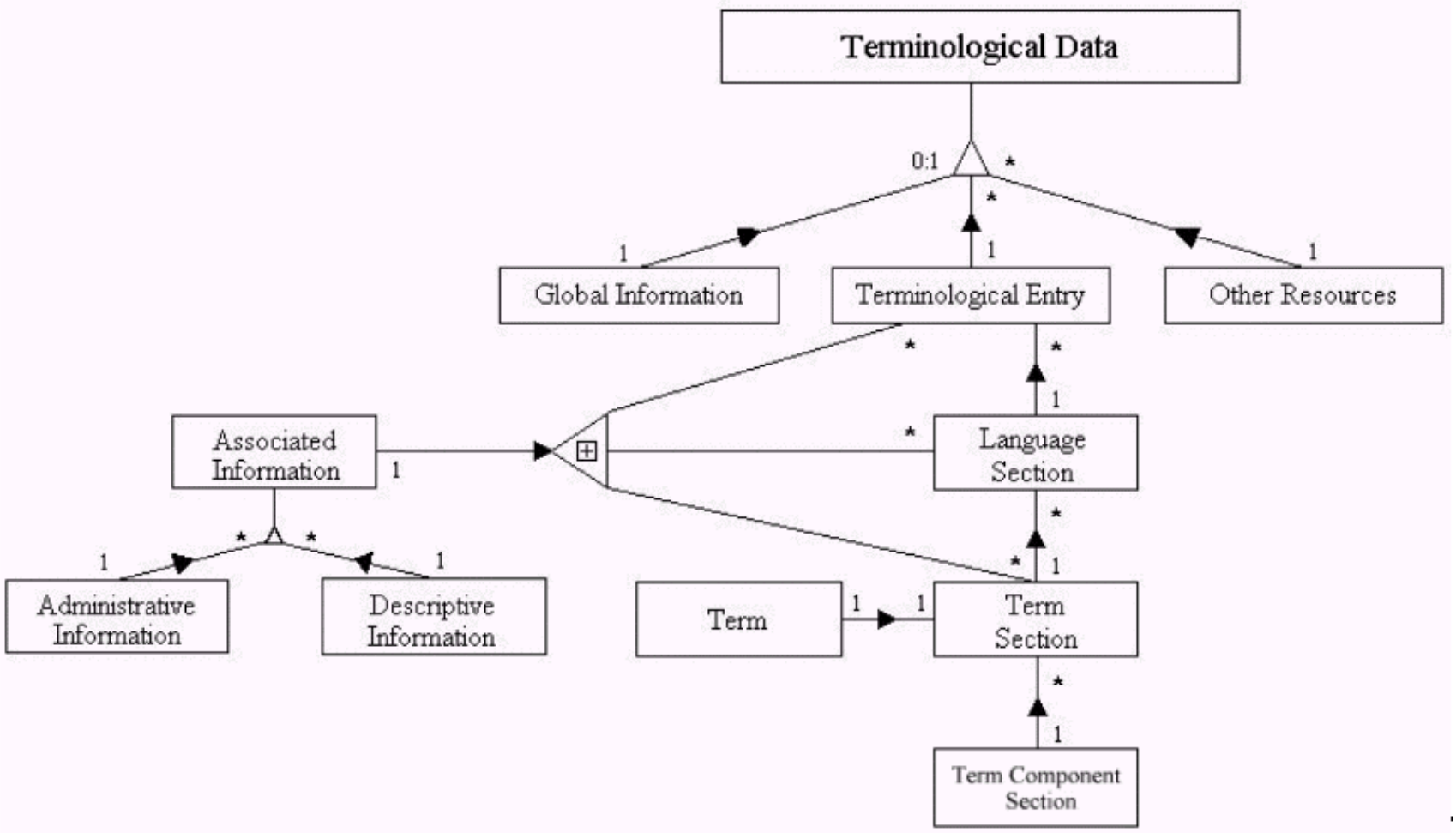
Terminology Representation Formats

- Multitude of formats for lexical/terminological data
  - E.g, MATER, TEI-lex/term, NTRF,OLID, MARTIF, TBX, IIF, TRANSTERM, GENETER
- MARTIF negotiated (ISO 12200)
- MARTIF specified MSC (ISO CD 16503; frozen)
- GENETER (ISO CD 17241; frozen)
- Meta-Model / TMF (ISO DIS 16642)

TMF / XLT / TBX

- Terminology Markup Framework (TMF)
  - Defines a framework for terminology interchange / representation languages (TMLs): Meta Model, DCS-Editor, Style
  - TMLs: Unicode, XML-based format
- Family of formats for lexical and terminological data (XLT)
  - Subset of ISO 12620 datcats, specific style (MARTIF-like
  - DXLT (Default XLT) = TBX of LISA

Meta Model


Terminological Data
0:1
*
*
1
1
1
Global Information
Terminological Entry
Other Resources
*
*
1
Associated Information
1
*
Language Section
*
1
*
*
1
1
Administrative Information
Descriptive Information
Term
1
1
Term Section
*
1
Term Component Section

ISO-TMS

- ISO 704
- Wright/Budin “Handbook of Terminology Management, Vol I and II”
- Concept-oriented approach of terminology management

Term autonomy principle

- Each term representing the (same) concept could be managed autonomously
- Could be Documented necessary term-related data categories

Data categories

- ISO 10241 “Preparation and layout of international terminology standards”
- ISO 12620 “Computer Applications in Terminology – Data Categories”

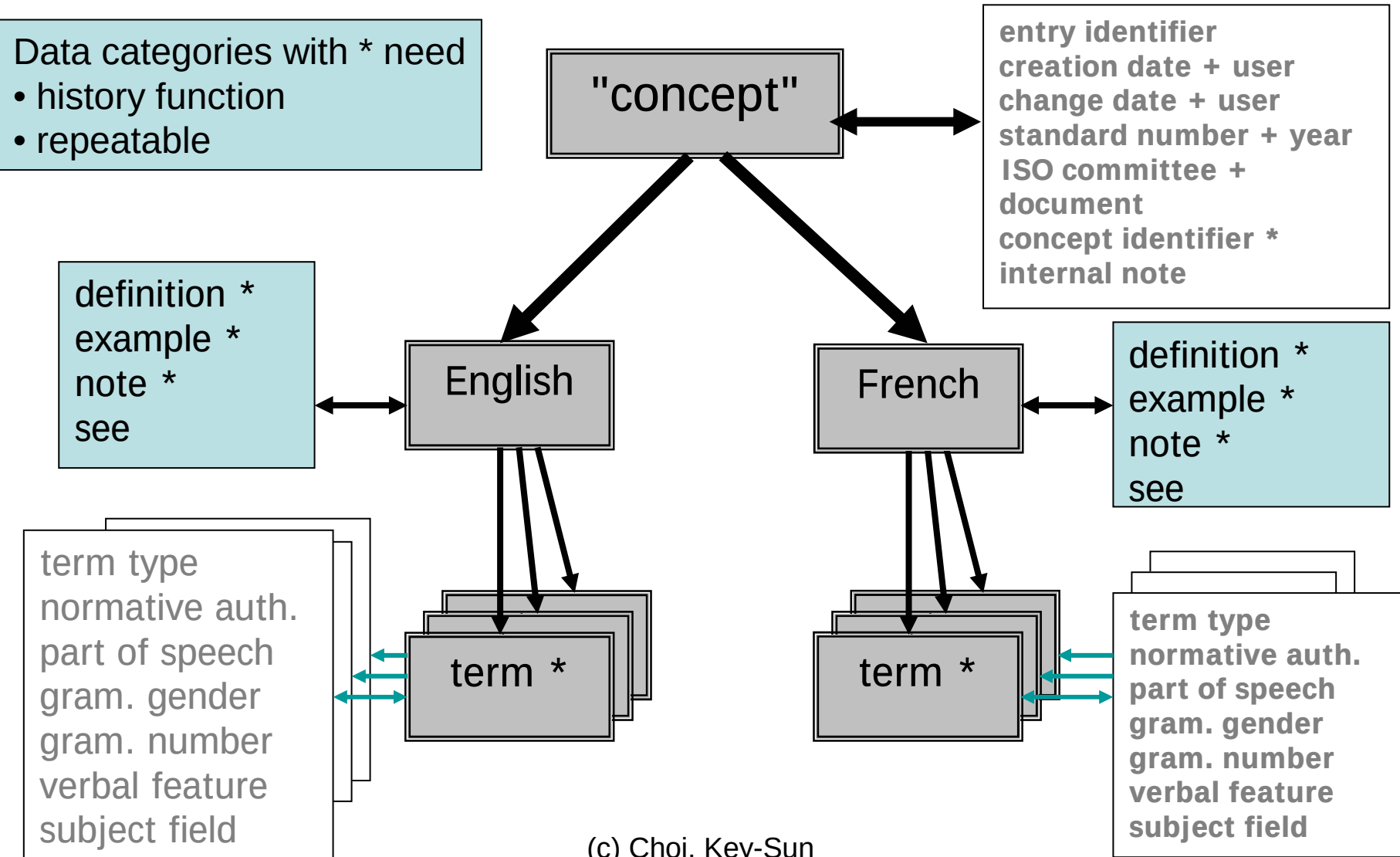
Templates

- For producing working drafts and the final standard directly from the ISO-TMS are necessary.

Experts' discussions

- Maintenance of provisional and intermediate results of the experts' discussions and decisions is of great importance.
- Work "history" maintenance
  - To trace and backtracking previous decisions and results

ISO-TMS database design



Extensions from TC 37/SC 3 to SC 4

- To develop a standard for data categories used in typical “SC 4 applications”
- May-be with different parts for different types of language resources (NLP lexica, texts, speech, etc.)
- May-be with different meta-models for different types of language resources
- Method: “learn from TMF history”

References

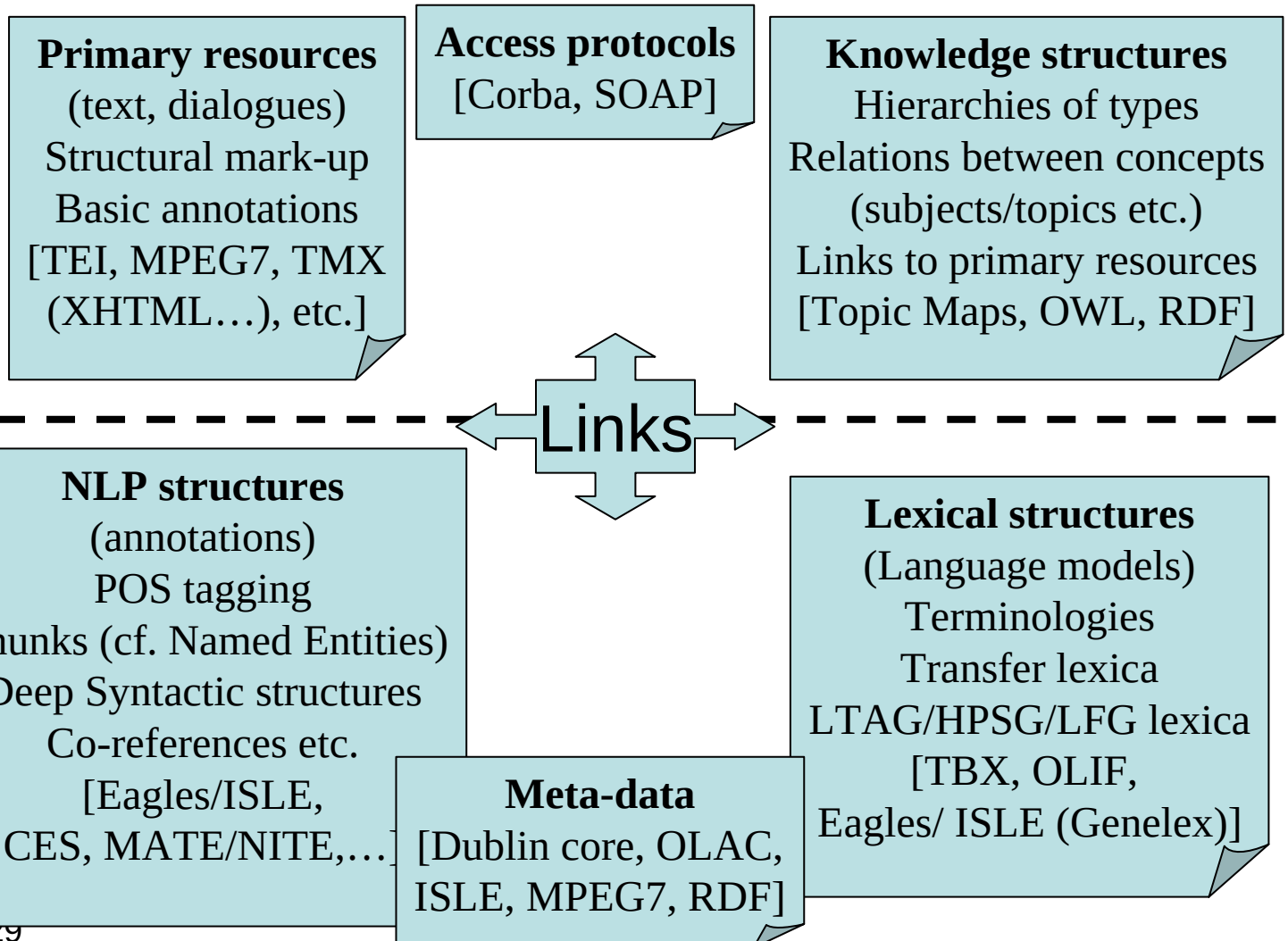
- MARTIF, ISO 12620 Data Categories, MSC, Meta-Model, TBX: www.ttt.org
- SALT Project: www.loria.fr/projets/SALT

ISO TC 37 / SC4 Language Resource Management

An overview

Laurent Romary

Standards for language processing



10 years ahead…

- Training a stochastic parser
  - And evaluating it…
- Putting together an information extraction system
  - And evaluating it…
- Parameterizing an MT system
  - And…
- Putting together a man-machine dialogue system in the context of an EU/NSF/national etc. project (e.g. Smartkom, MIAMM, …)

Context

- **ISO TC37 -** Terminology and other language resources
  - **SC3 -** Computer applications in terminology
    - ISO 12200 - Martif
      - Latest version of TEI Terminology chapter
    - ISO 12620 - Data categories (under revision)
    - ISO DIS 16642 - TMF (Terminological Markup Framework)
  - **SC4 -** Language resources

Goals of ISO / TC 37/SC 4

- prepare international standards/guidelines for effective language resource management in mono- and multi-lingual applications
- develop principles and methods for creating, coding, processing and managing language resources
  - written corpora, lexical databases, spoken language corpora, etc.
- Focus :
  - *data modeling* (à la MPEG7?)
  - data exchange, evaluation

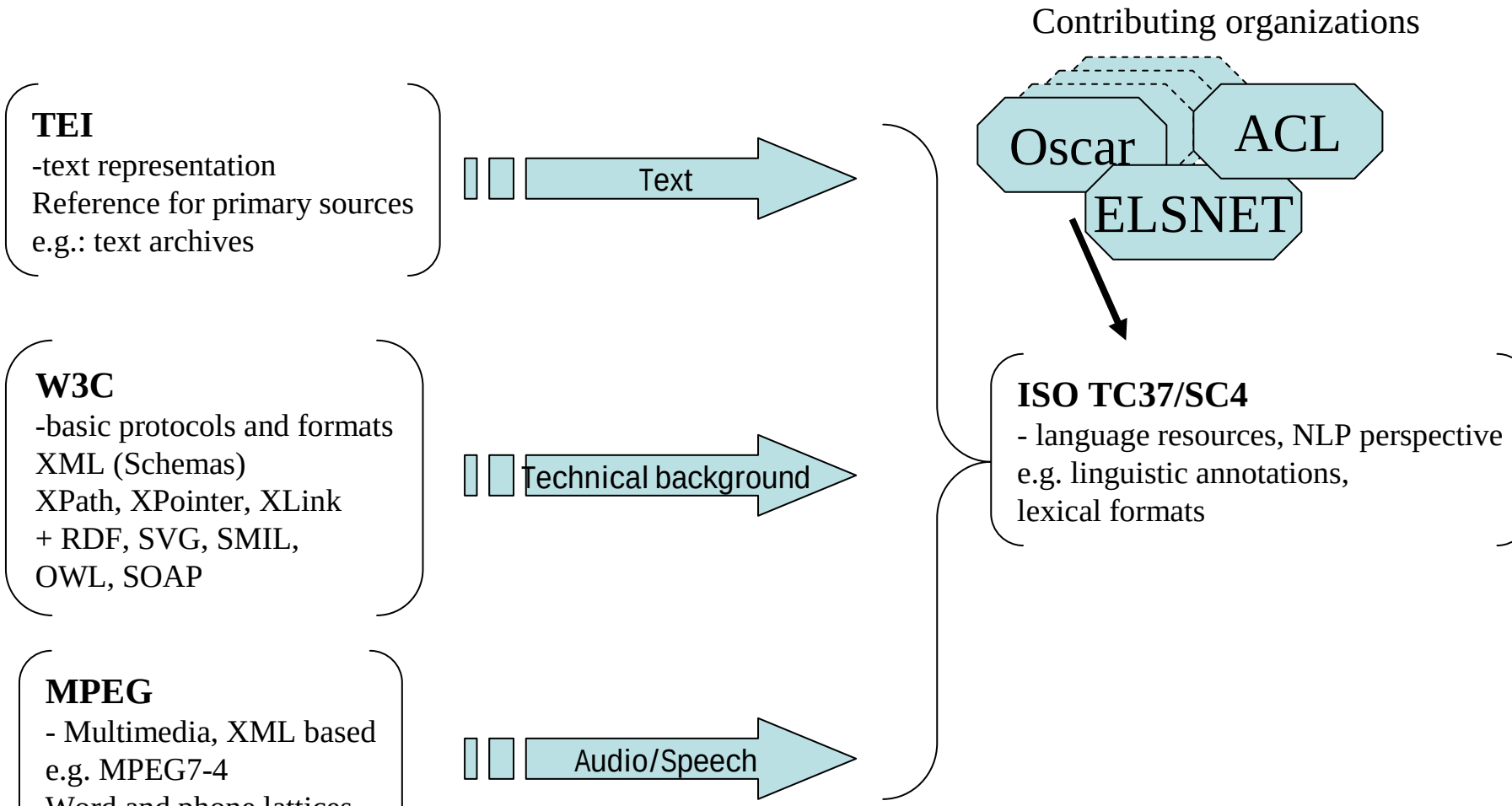
TC37/SC4 details

- Scope
  - Platform for designing and implementing linguistic resource formats and processes
    - Multi-layer annotation of linguistic resources
    - Exchange of information between NLP modules
- General strategy
  - Involve a wide community from academia and industry
    - through national standardizing bodies
- Agenda
  - Constituency meeting and technical workshop at LREC (May 2002)
  - Current:
    - identification of possible work items and working groups
    - Implication of new experts and work agenda planning

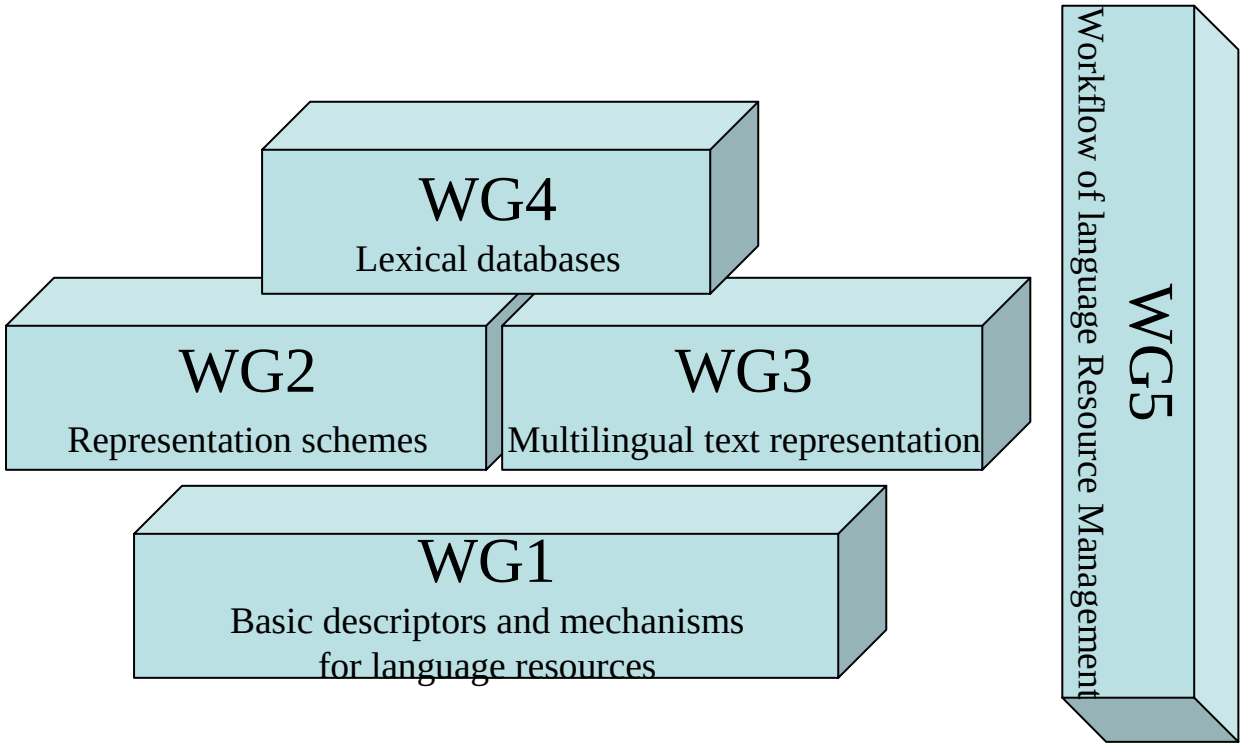
Organization

- Chair:
  - Laurent Romary, France
- Secretary:
  - Key-Sun Choi, Korea
- International Advisory Committee
  - Permanent Chair: Prof. Antonio Zampolli, Italy

SC4 and other standardizing bodies



TC37/SC4 overall rationale



TC37/SC4 Work Items

- WG1
  - PWI: Terminology of Language Resources
  - PWI: Linguistic annotation framework
  - PWI: Meta-data for multimodal and multilingual information
- WG2
  - PWI: Structural content representation scheme
  - PWI: Multimodal content representation scheme
  - PWI: Discourse level representation scheme

TC37/SC4 Work Items - cont.

- WG3
  - PWI: Translation Memory, Alignment of parallel corpora
  - PWI: Segmentation and counting algorithms (characters, words, sentences etc.)
  - PWI: Meta-markup for GIL (Globalization, Internationalization and Localization)
- WG4
  - PWI: NLP Lexica
- WG5
  - PWI: Validation of language resources
  - PWI: Net-based distributed cooperative work for the creation of language resources

Report on WG1 activities

- WI: Terminology of language resources
  - Basic requirements - K.-D. Schmitz
  - Existing sources for bootstrapping a TermBank in SC4 - K.-S. Choi (proj. leader)
- WI: Linguistic Annotation Framework
  - data categories for lang. res. - S.E.Wright
  - Basic requirements and workplan - N. Ide (proj. leader)
- TF: Meta-data for multimodal and multilingual information
  - Task description and scope - P. Wittemburg (rep. By L. Romary)

Report on WG1 activities (cont.)

- Related activities
  - Semantic content representation - K.-Y. Lee
  - Dialogue architectures - L. Romary
  - …

WG1 - PWI

- Terminology of Language Resources
  - Basic terminology of the various sub-fields of language resources and general methodology
  - Possible sources:
    - ISO 1087
    - LREC proceedings + KAIST
    - Support from GTW

WG1 - PWI

- Linguistic annotation framework
  - Basic mechanisms and data structures for linguistic annotation and representation [data architecture]
    - Methods and principles for the design of an annotation scheme
    - Structural nodes and information units, Data category specification
    - Linking and pointing mechanisms, Feature Structures, Meta-Markup
    - « Stand-off » and « in-line » views - equivalences, combining levels.
    - Administrative data categories
  - Possible sources:
    - TMF, iso12620-revised, Mate/NITE (general methodology)
    - TEI (Linking mechanisms, feature structures)
    - Link with *Linguistic DS*

WG1 - PWI

- Meta-data for multimodal and multilingual information
  - Description of a meta-data representation scheme to document linguistic information structures and processes
    - General content description
    - Local content description
  - Possible sources:
    - OLAC, Mile, TEI Header
  - Liaison: TC46 (SC9), MPEG7/MDS, SCORM

WG2 - PWI

- Structural content representation scheme
  - Definition of annotation/representation scheme(s) for morpho-syntax and syntax, to be used for annotation and interchange purposes
    - Meta-model for morpho-syntactic annotation
    - Meta-model(s) for syntactic annotation (lexicalized grammar, elementary trees, dependancy structures)
    - + corresponding Data category registries
  - Possible sources:
    - Eagles, TAGML, Linguistic DS
    - SIGPARSE

WG2 - PWI

- Multimodal meaning representation scheme
  - Representation scheme for the semantic content of multimodal information (textual, spoken, graphical and gestural)
    - Meta-modal for content representation (Events, participants, etc.)
    - Data category registry for multimodal content
  - Possible sources:
    - SIGSEM working group on semantic content
      - Chair: Harry Bunt
  - « Liaison »
    - Semantic web activities

WG2 - PWI

- Discourse level representation scheme
  - Meta-model for discourse and dialogue representation
  - Meta-model for discourse level annotation (e.g. reference annotation)
  - + corresponding DatCat registry
  - Possible sources:
    - SIGDIAL
    - DRI - Discourse Resource Initiative
    - Mate

WG3 - PWI

- Translation Memory, Alignment of parallel corpora
  - Provides formats for the representation of multilingual textual data as produced in translation activities or constructed from existing primary sources
  - Sources:
    - OSCAR/TMX for translation memories
    - TEI based linking mechanism (or see WI-1) for Parallel texts

WG3 - PWI

- Segmentation and counting algorithms (characters, words, sentences etc.)
  - Provide methods for segmenting streams of text with markup and means to for counting the corresponding segments
  - Possible sources:
    - OSCAR

WG3 - PWI

- Meta-markup for GIL (Globalization, Internationalization and Localization)
  - Identification of the specific markup modules needed to perform GIL activities
  - Possible sources:
    - OSCAR/OpenTag

WG4 - PWI

- NLP lexica
  - Lexicon representation formats for the various types of NLP applications (Machine Readable Lexica)
    - Define a set of meta-models (classes of applications)
    - Specific data categories (derivation, phonology, etc.)
    - Based on the work done in other work items
  - Possible sources
    - Eagles, Multext, ISLE Computational lexicon Working group, OLIF

WG5 - PWI

- Validation of language resources
  - Defines guidelines and requirements for producing and distributing high quality language resources
  - Contacts:
    - ELRA, TEI
  - Possibles sources:
    - To be defined

WG5 - PWI

- Net-based distributed cooperative work for the creation of LRs
  - Principles and methods for designing collaborative and cooperative compilation of LRs
  - Define what is specific to LRs with regards
    - Tracability of resources, version control, validation, quality management
    - Protocols (Corba, SOAP), Workflow standards, Data management
  - Sources: To be defined

WG 1-2

Descriptors and Mechanisms for Language Resources

Task Description and Scope

Peter Wittenburg, Daan Broeder

Focus

- Overview about existing projects/initiatives and monitor its usage
- Linkage with activities of the emerging Semantic Web

What to do for WG1-2

- Determine the
  - Scope of language resources
  - Needs of the community
  - Existing initiatives relevant to the language resource domain
- Develop a scenario how metadata will be used in the Semantic Web
- Determine the set of descriptors and their vocabularies useful to describe language resources
- Define
  - All relevant terminological units (concepts and terms in major languages) and their relations
  - Suitable frameworks for the definitions

Requirements for manageable domain of language resources

- Unique identifier for all resources
  - conforming to the web standards (URI)
- Metadata descriptions to facilitate management
  - Major characteristics for large spectrum of different types of language resources
  - Virtual management domains:
    - Accessible metadata to prohibited due to commercial, ethical or legal reasons
  - Interoperability mechanisms
    - Syntax, semantics (data categories, vocabularies, relations)
- Description level for Human + program readability
- Complete workflow cycle for “management”
  - Process of resource creation, enhancement, integration, discovery, exploitation, archiving and deletion
- Web and multilinguality

Scope of language resources

- Electronically available resources
- Written, spoken and non-verbal (gesture, sign, facial expression and other modalities)
- Examples
  - Websites which contain language
  - Publications (books)
  - Recordings of sign language
  - Lexica
  - Multimedia recordings or multimedia extensions
  - Mono- and multilingual

View on language resources

- Examples of language resources
  - Technical documents of cars
  - Material part of learning objects
  - Annotated film movies
- Completely different communities: e.g.,
  - Technical documentation engineers
    - Descriptors to easily retrieve a relevant document
  - Linguist
    - For research perspective, other type of descriptors
- Description of content

Different sets of document management

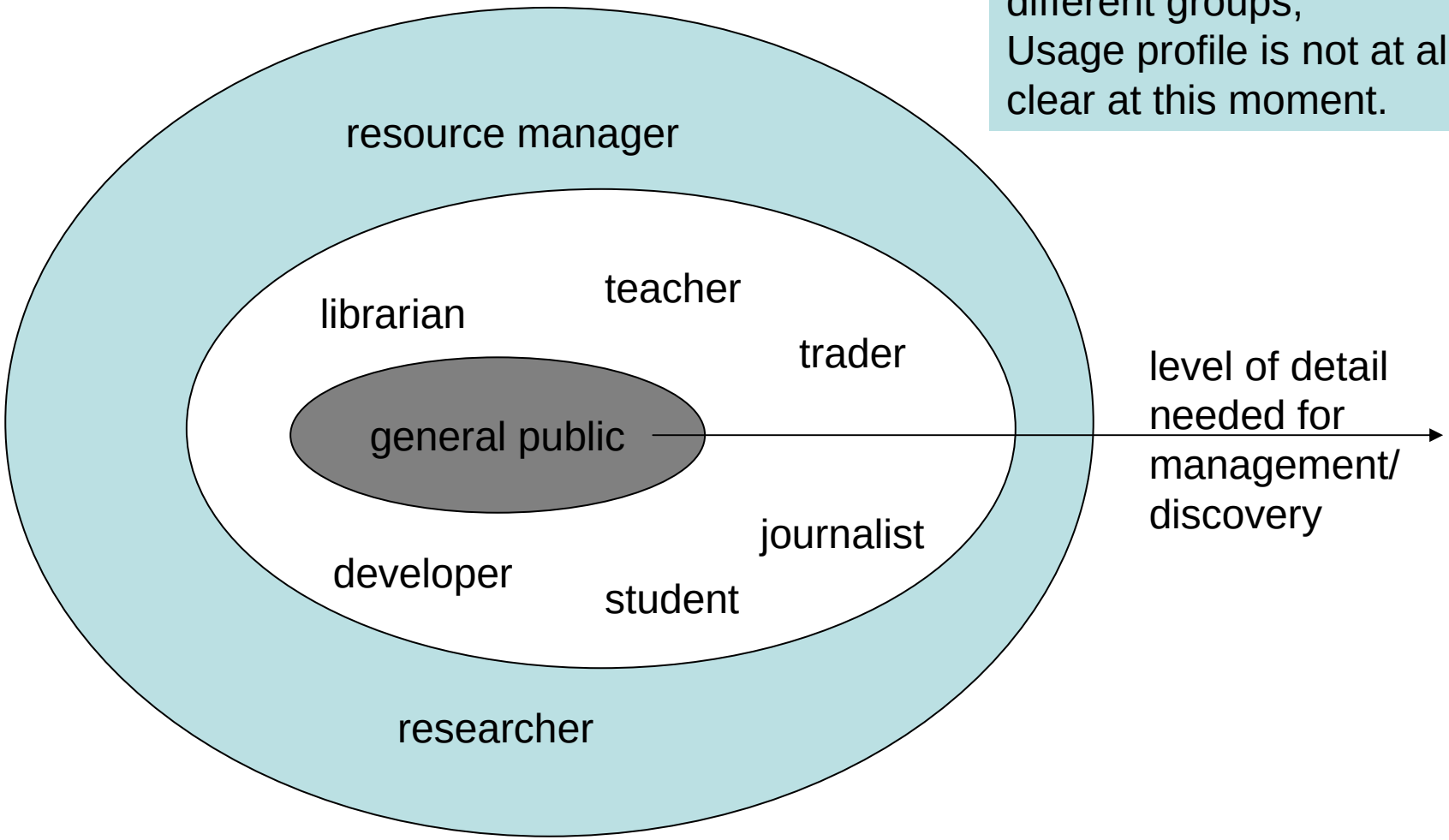
- Differences
  - Dublin Core
  - CEI/IEC 82045-1
- Scope restriction in SC4/WG1-2
  - “Language Resources” to the study of language
  - Metadata descriptions describing language resources
    - are not meant to be language resources themselves in the context of this note.
    - Solely used for resource management and discovery purpose
    - Does not exclude that they will be viewed as LRs within the context of other work.

LR Community

- Usage scenarios for LR
  - General public – general information on many subjects
  - Resource traders, researchers or language engineers
    - Selling LR, deriving new grammar, calculating the parameters of statistical recognition algorithms

Levels of abstraction: LR community

2nd level includes many different groups;
Usage profile is not at all clear at this moment.



Strongly related metadata initiatives

- Media and film community: MPEG7
  - Started by looking at existing standards such as SMPTE and the emerging requirements of the community.
  - Exhaustive element set combining
    - Suggestion for annotating film productions
    - Creating metadata descriptions
- Focus in film industry
  - To support the production process
    - Annotate movies with low-level features such as “scene change”

Strongly related metadata initiatives: MPEG7

- MPEG7 intention
  - In the object oriented MPEG4 decoding scenario
  - To support the query, selection and filtering process of the user
    - Easily to assemble clips and other information
    - To new personalized presentations.
- MPEG7 Description Definition Language
  - To define Descriptors and Description Schemes that is based on XML.
  - Many descriptors and description schemes – already been defined including a “linguistic” one defining how linguistic phenomena can be encoded

Strongly related metadata initiatives: MPEG7 -- DC

- Harmony project
  - Very restrictive mapping of MPEG7 metadata elements to Dublin Core specifically to not extend the semantics of DCMES
- MPEG7 is not particularly designed for the view on multimedia language resources people from the linguistic community have, but MPEG7 will have strong impacts

Strongly related metadata initiatives: IEEE - LOM

- LOM (Learning Object Metadata)
  - Start from DCMI
  - Propose an exhaustive set of elements
    - For the sufficient description of learning objects
  - … structure
    - specifies the set of elements together
      - With constraints for
        - » Order, value range, basic data type
    - groups the elements into categories
  - To include aggregate and simple elements
    - Similar to IMDI implicit structure
  - To define controlled vocabularies for a number of data elements.

Strongly related metadata initiatives: CEI/IEC 82045-1

- In the area of content management
- Joint effort (JWG15) for a metadata element set from ISO TC 10 and IEC SC3B
- Proposal statement
  - Management data as data about the content of an electronic or paper document, necessary to manage it in an Electronic Document Management System
- Broad analysis of possible documents and document collections including their various types of relationships during their life-time, an exhaustive metadata set is developed.
- Many elements are grouped together in a hierarchy for intelligibility reasons.

Construction of WI-2: Relevant initiatives

- TEI: as initiative having worked extensively in defining structures of textual resources
- DC: as most important metadata initiative world-wide with a claim for general coverage and interoperability
- OLAC: as the DC-based initiative in the domain of language resources
- IMDI: as the initiative in the domain of language resources covering more detailed descriptions
- MPEG7: as a highly relevant initiative in a closely related domain

Construction of WI-2: Relevant initiatives

- IMS/LOM: also as a highly relevant initiative in a closely related domain
- IEC 82045-1: also as a highly relevant initiative in a closely related domain
- COLLATE: as an initiative gathering much data in the domain of language resources almost ready to define formal metadata sets
- Terminology Initiatives: as initiatives which know much about the definition of data categories
- RDF/W3C: as an initiative which has the experts to show the way from metadata to the Semantic Web

Experts in metadata requirements

- Corpus and field linguistics
- Language engineering
  - Text-based work
  - Multimodal work
- Artificial intelligence
- Phonetics
- Psycholinguistics

Multimodal Semantic Representation

Harry Bunt, Laurent Romary, Kiyong Lee,

http://let.kub.nl/people/bunt

Meaning representation from different perspectives

- Language (text and speech) interpretation
- Language generation
- Design of embodied agents
- Interpretation of gestures and facial expressions
- Fusion and coordination of modalities
- Multimodal dialogue system architecture

What should be represented in multimodal semantics

- Semantic and pragmatic aspects of meaning
- Information in gestures, postures, facial expressions
- Nonlinguistic information in speech and text
- Metadata: time stamps, processing information, …
- Social and cultural information
- Links to other information

Basic components and mechanisms in multimodal semantics

- Representational constructs
- Mechanisms for linking and sharing
- Specific representational issues, such as
  - Underspecification techniques
  - Flatness of representations
- Which representation formalisms should be analyzed for best practices, lessons learned, common underlying ideas, special features, …

Relation to neighboring activities: multimodal semantics

- XML, RDF, semantic Web
- Ontologies, domain modeling, user modeling
- Lexicon formats
- Annotation schemes

Background organization for WG2

http://www.sigsem.org/

http://cwis.kub.nl/~fdl/research/ti/sigsem/iwcs/iwcs5

ISO/TC 37/SC 4/WG1-1

Language Resource Management

Linguistic Annotation Framework

21-23/Nov/2003

Pont-à-Mousson, France

<!-- Morphosyntactic level -->

```
<struct type= “W-level”
   <seg target = “#w3”>
   <feat type=“lemma”>de</feat>
   <feat type=“pos”>PREP</feat>
</struct>
<W-level xsi:type=“struct” xlink:href=“#w3”>
   <lemma xsi:type=“Feat”>de</lemma>
   <pos xsi:type=“Feat”>PREP</pos>
</W-level>
```

Revised --

```
<struct iso-sc4:type= “W-level”
   <seg xlink:href = “#w3”>
   <feat type=“lemma”>de</feat>
   <feat type=“pos”>PREP</feat>
</struct>
<W-level xlink:href=“#w3”>
   <lemma>de</lemma>
   <pos>PREP</pos>
</W-level>
```

General scheme: (Nancy Ide)

- Expressive adequacy
  - Represent all varieties of linguistic information
  - Media independent
  - Rely on existing or developing standards for multi-media
- Semantic adequacy
  - Representation structures must have a formal semantics
  - Centralized way of sharing descriptors and information categories
  - Definitions of operations
- Incrementality
  - Support for various stages of input interpretation and output generation
  - Representation of partial/under-specified results and ambiguities, alternatives, etc. (Support for under-specification)
- Uniformity
  - Representations utilize same “building blocks” and the same methods for combining them
- Openness
  - Not dependent on a single linguistic theory
- Extensibility
  - Compatible with alternative methods for designing representation schemas
- Human readability

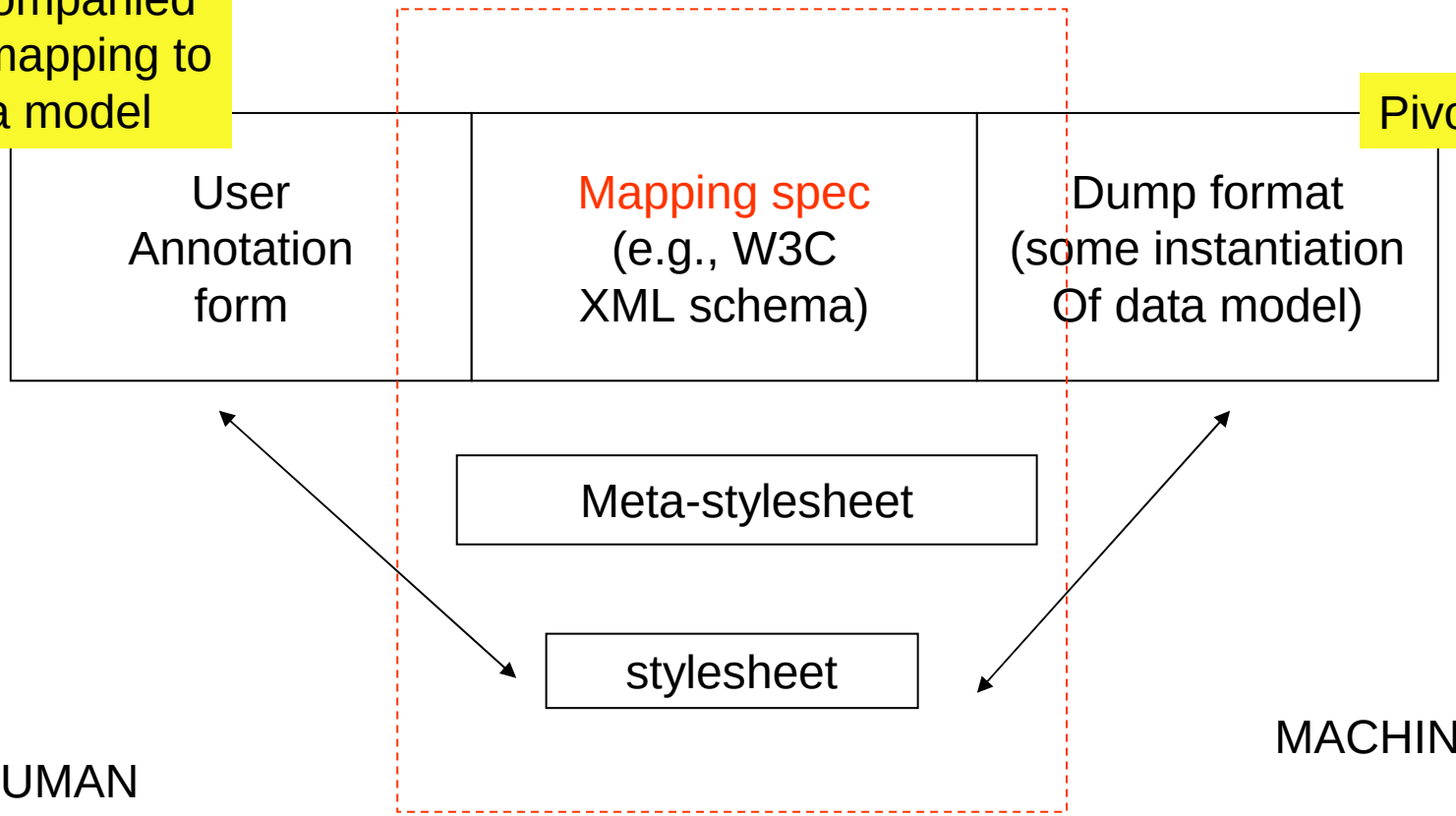
Principle of general scheme (discarded)

- Expressiveness: multi-media coverage
  - Don’t reinvent existing or developing standards
    - (e.g., MPEG7)
  - Media-independence

Relation between document design and data model

Not intended for archive

Accompanied By mapping to Data model



Or the functional equivalent thereof