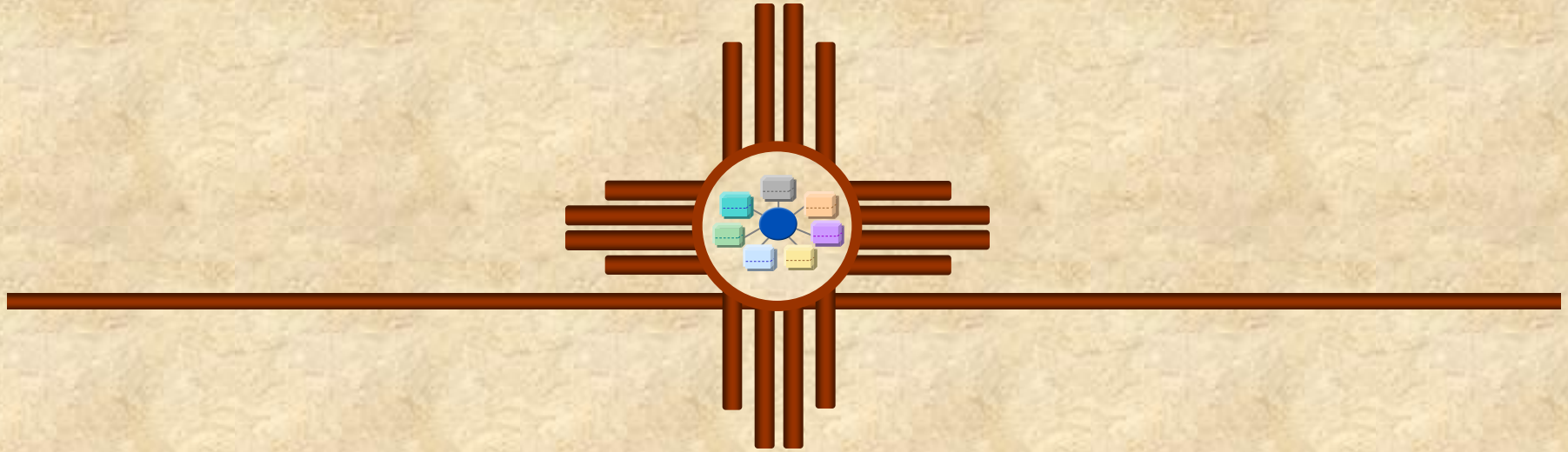
A Metamodel to Represent Terminology Data Collections

Laurent Romary

A Metamodel to Represent Terminology Data Collections



Open Forum 2003 on Metadata Registries

Terminology and Ontologies Track

20-24 January 2003

Laurent Romary

Laboratoire Loria-INRIA

Summary

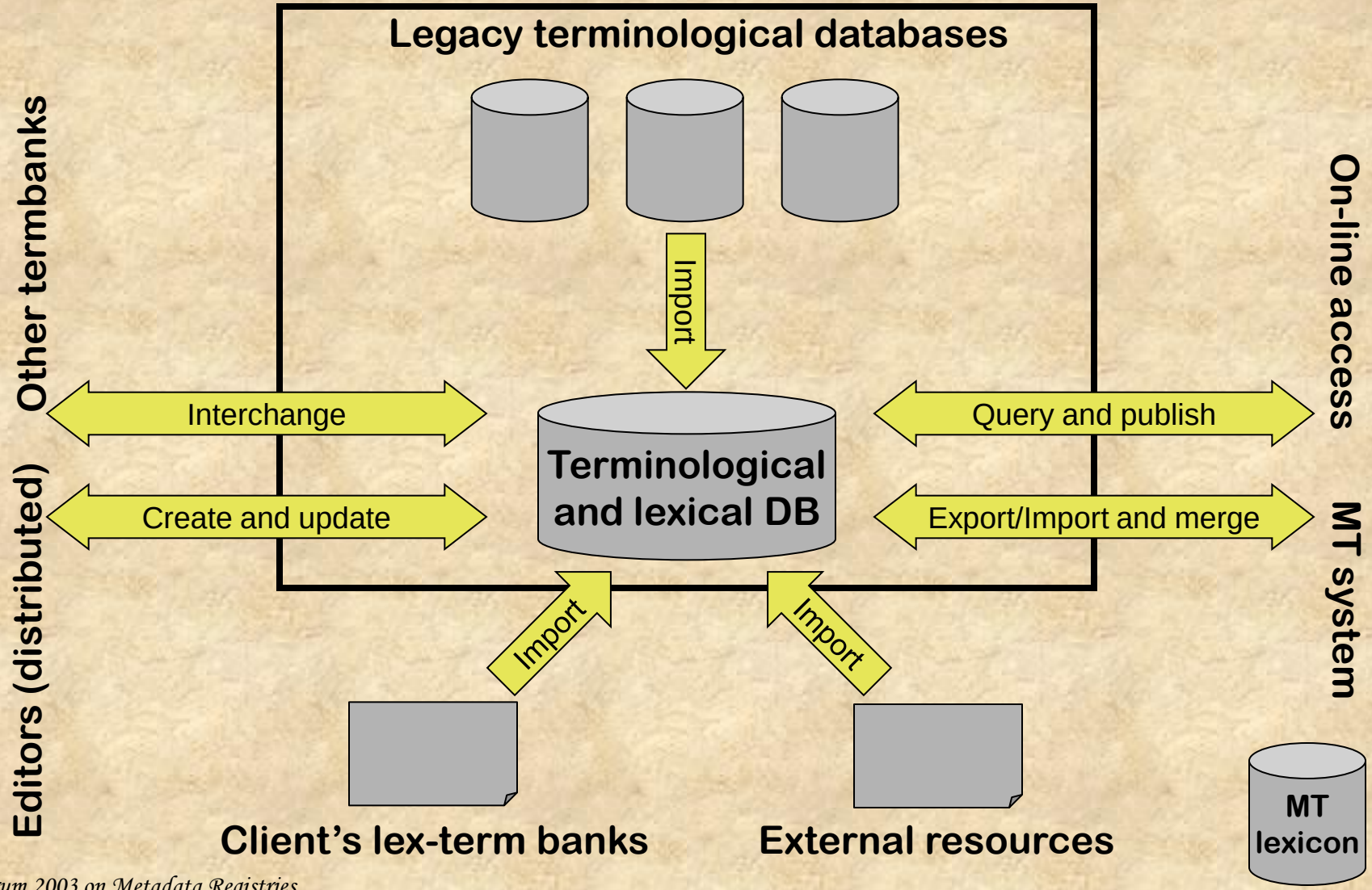
- From terminologies to ontologies (and back…)
  - Experience gained in TC37/SC3 while working on ISO 16642 (Terminological Mark-up Framework)
    - Abstracting away from XML structures
  - Paving the way for future work within ISO TC37/SC4
    - The central role played by the metadata registry
    - Relation between TC37/SC4, ISO 11179 and W3C

TC37/SC1: Principles and methods
TC37/SC2: Terminography and Lexicography
TC37/SC3: Computer applications for terminology
TC37/SC4: Language resource management

General context

- Designing a platform for representing terminological data
  - ISO TC37/SC3 context (computer applications in terminology)
    - Competition between two formats (i.e. two DTDs)
    - Design of ISO 16642: *TMF - Terminological Markup Framework*
  - European IST/Salt project
    - Working on the interoperability of lex-term formats

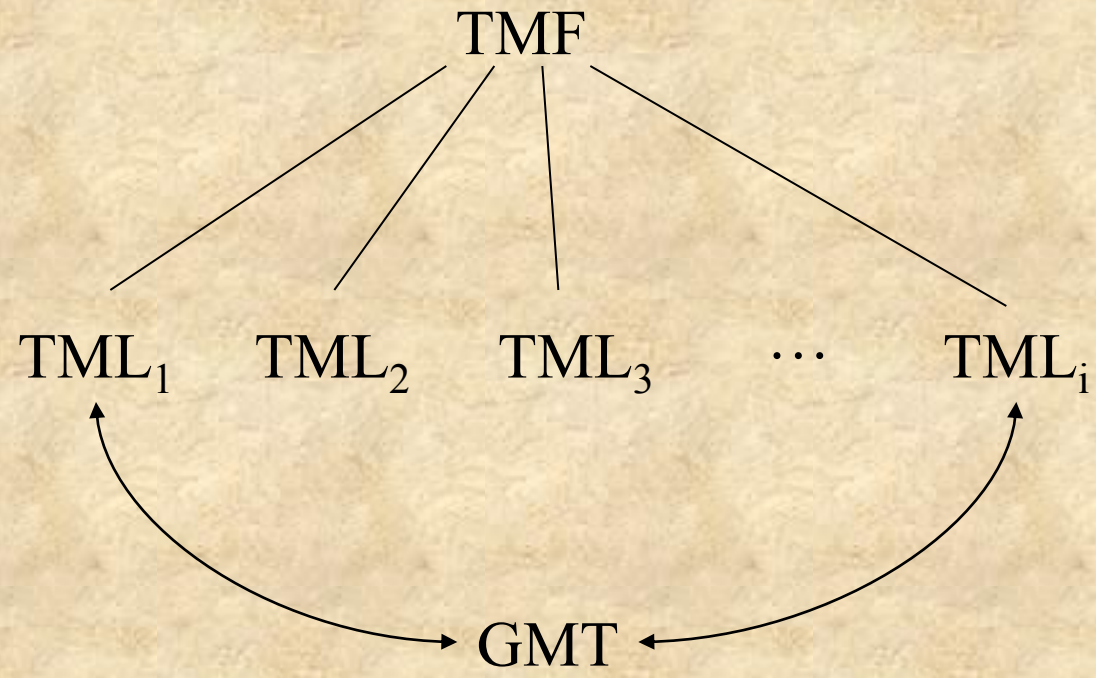
The ecology of lex-term data



Objectives of ISO 16642

- Providing a platform to:
  - Describe existing data structures
    - How does a client's information relate to one's own terminological database
  - Design company specific environments
    - E.g. to integrate lexicographic information related to MT
  - Identify ways of mapping these structures to industrial standards
    - E.g. export data in TBX

A family of formats

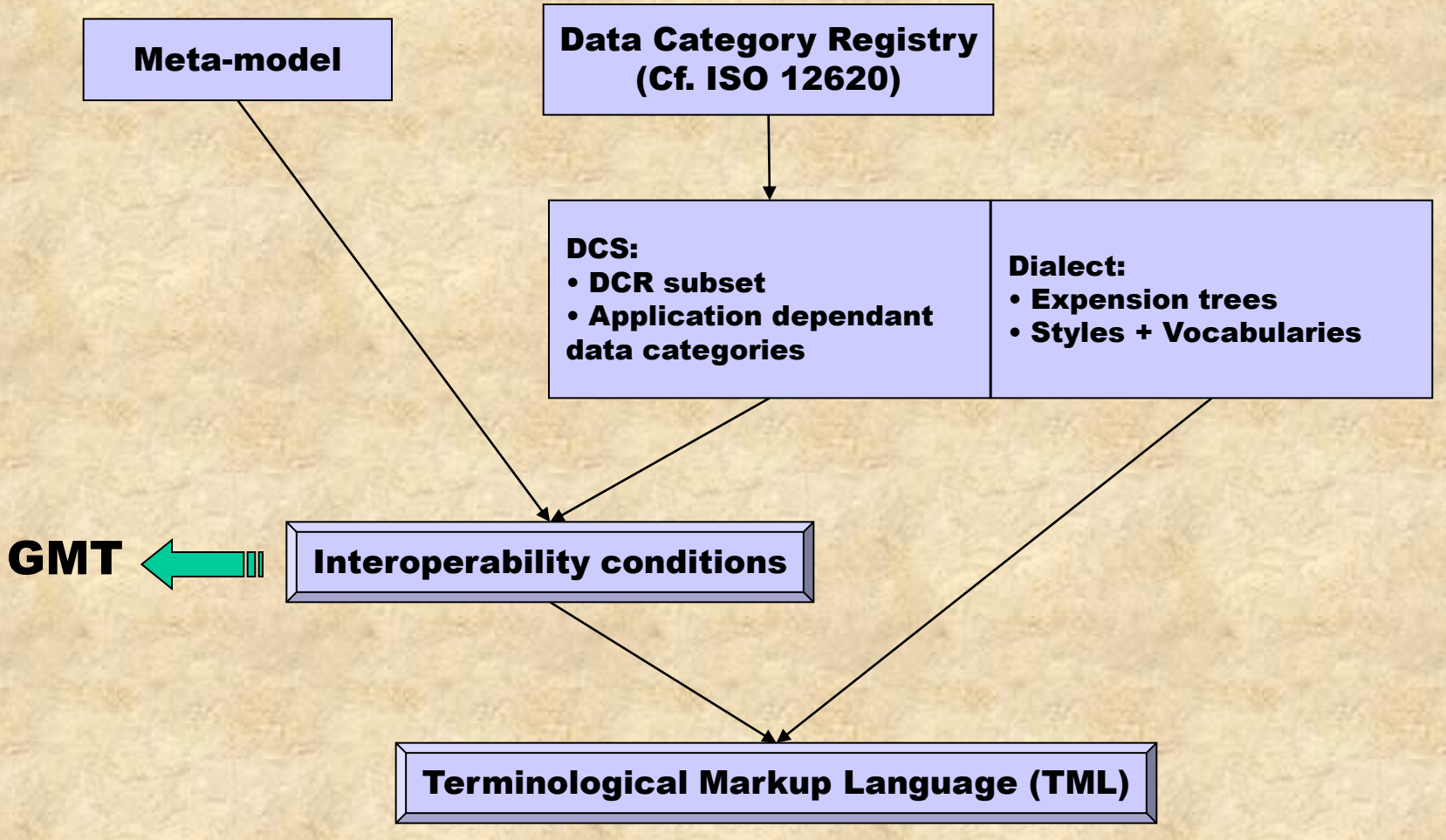


TMF - Terminological Markup Framework
TML - Terminological Markup Language
GMT - Generic Mapping Tool

General principles

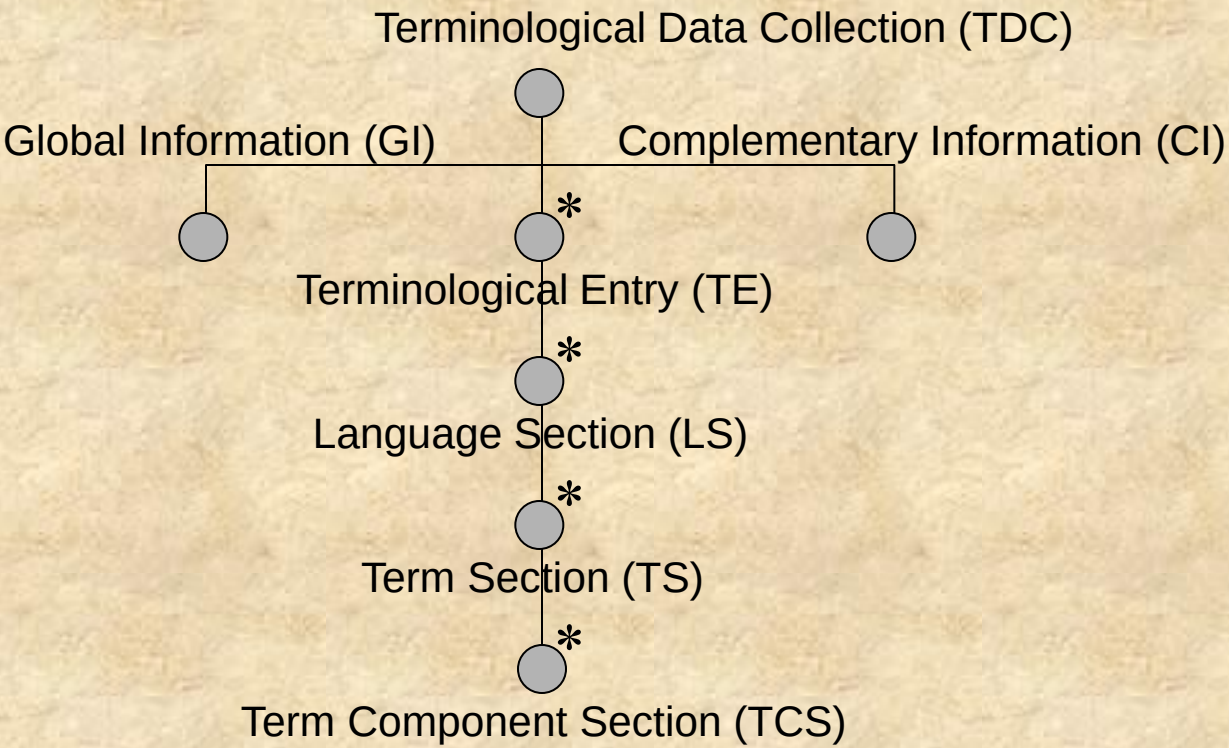
- Expressing constraints for representing computerized terminologies
  - What is the underlying structure of computerized terminologies?
  - Which data categories are used and under what conditions?
- Maintaining interoperability between representations
  - Providing a conceptual tool for comparing two given formats

Designing a TML



DCR - Data Category Registry
DCS - Data Category Selection
GMT - Generic Mapping Tool

Meta-model


Terminological Data Collection (TDC)
Global Information (GI)
Complementary Information (CI)
*
Terminological Entry (TE)
*
Language Section (LS)
*
Term Section (TS)
*
Term Component Section (TCS)

Data categories

- Existing background:
  - ISO 12620: *Computer applications for terminology - data categories*
  - Around 300 entries:
    - Term, Part of speech, Preferred term, Animacy (Animate, Inanimate)
    - Abbreviated form for, Broader concept generic, …
- Towards a formal description of data categories:
  - RDF model of data category
    - Editing, on-line browsing, TML modeling
  - Basic attributes (inspired by ISO 11179)
    - Identification of the data category (ID, name, definition etc.)
    - Values (Character data, Integer, picklist etc.)
    - Locations of the data category in relation to the meta-model
    - Administrative fields to maintain one's own specification

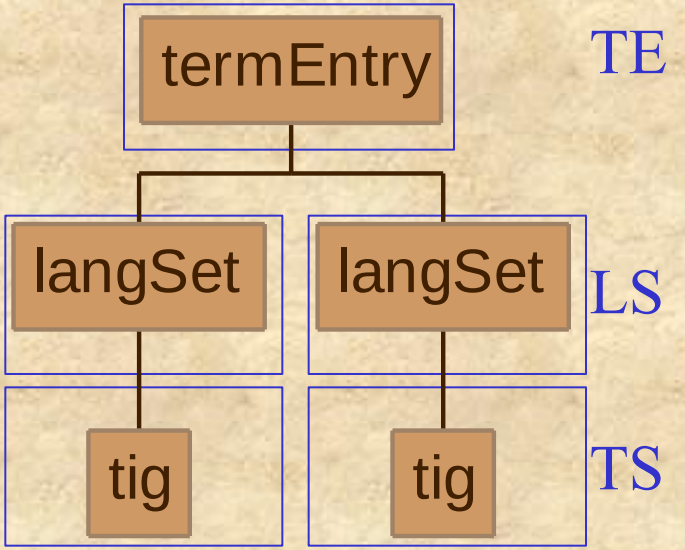
Putting 16642 at work: decomposition of a a terminological entry

TBX representation

```
<termEntry id='ID67'>
    <descrip type='subjectField'>manufacturing</descrip>
    <descrip type='definition'>A value between 0 and 1 used in
    ...</descrip>
    <langSet lang='en'>
        <tig>
            <term>alpha smoothing factor</term>
            <termNote type='termType'>fullForm</termNote>
        </tig>
    </langSet>
    <langSet lang='hu'>
        <tig>
            <term>Alfa ...</term>
        </tig>
    </langSet>
</termEntry>
```

Identifying the structural skeleton

id='ID67' [attribute]
subjectField=' manufacturing ' [typedElement]
definition='A value...' [typedElement]



lang=' en ' [attribute]

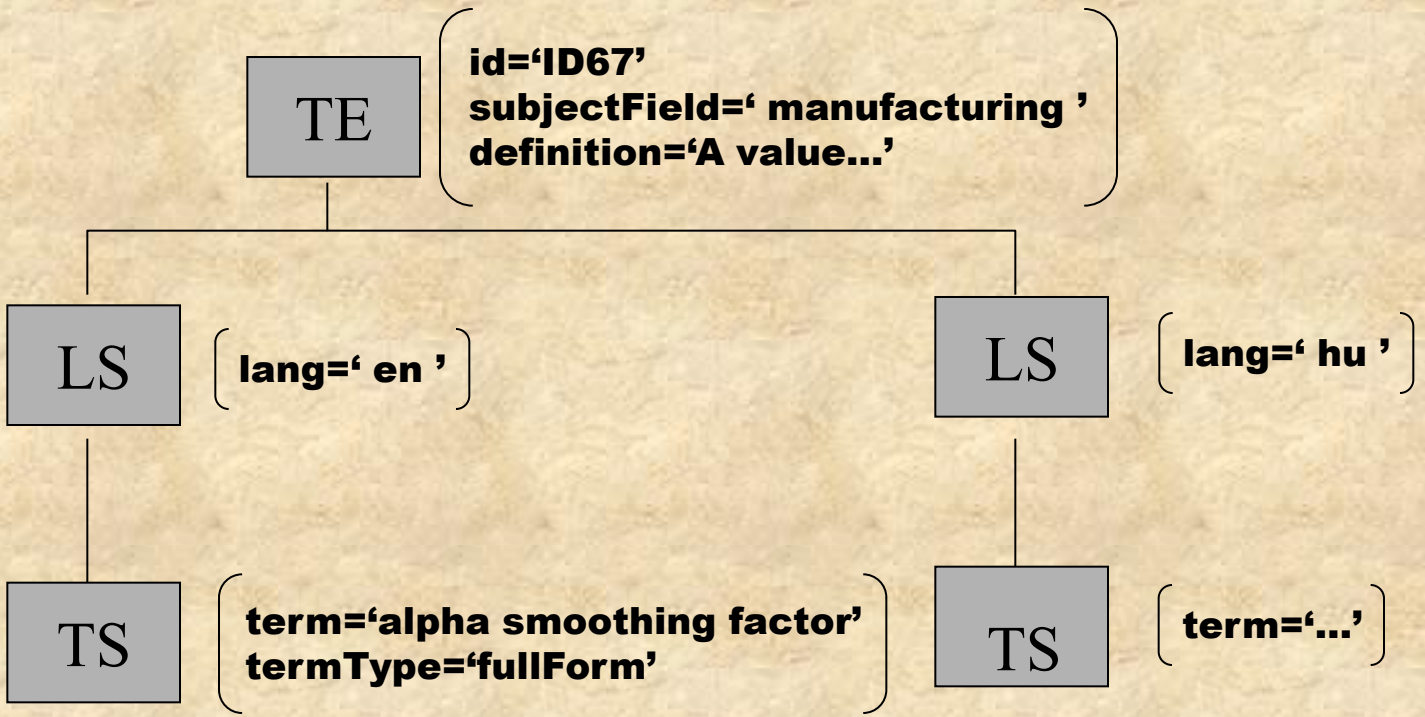
lang=' hu ' [attribute]

term='...' [element]

term='alpha smoothing factor' [element]
termType='fullForm' [typedElement]

TE - Terminological Entry
LS - Language Section
TS - Term Section

TMF information model



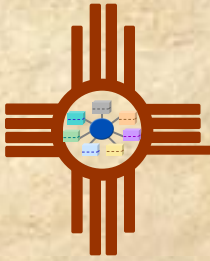
GMT representation

```
<struct type= “TE” >
    <feat type= “id” >ID67</feat>
    <feat type= “subjectField” >manufacturing</feat>
    <feat type= “definition” >A value between 0 and 1 used in ...</feat>
    <struct type= “LS” >
        <feat type= “lang” >en</feat>
        <struct type= “TS” >
            <feat type= “term” >alpha smoothing factor</feat>
            <feat type= “termType” >fullForm</feat>
        </struct>
    </struct>
    <struct type= “LS” >
        <feat type= “lang” >hu</feat>
        <struct type= “TS” >
            <feat type= “term” >Alfa ...</feat>
        </struct>
    </struct>
</struct>
```

Styles and vocabularies

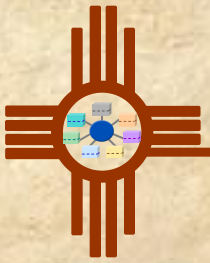
Implementing a DatCat

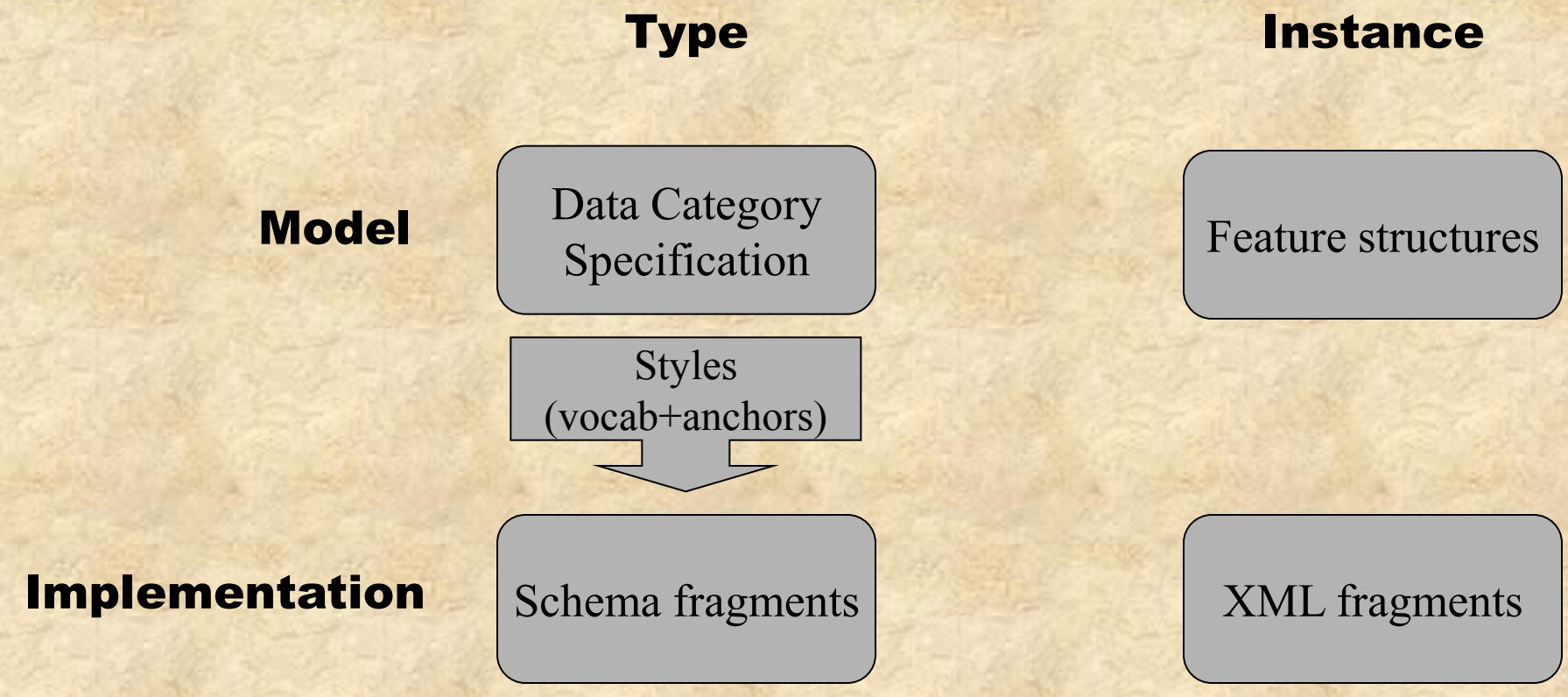
- Definitions:
  - ‘ style ’ — The way a given DatCat is implemented as an XML object
  - ‘ vocabulary ’ — symbols needed to express the implementation of a given DatCat in its associated style
- E.g.:
  - DatCat: /definition/
  - Style = Element
  - Vocabulary = [“def”]
  - <def>pencil whose casing …</def>



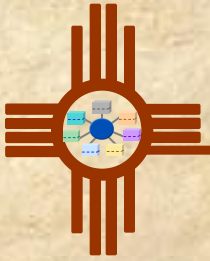
From an information model point of view…

Modeling Information Units





Modeling Structure



	Type	Instance
Model	Meta-Model (Fixed by 16642)	Structural skeleton
	Expansion trees	
Implementation	XML Schema fragments	XML outline

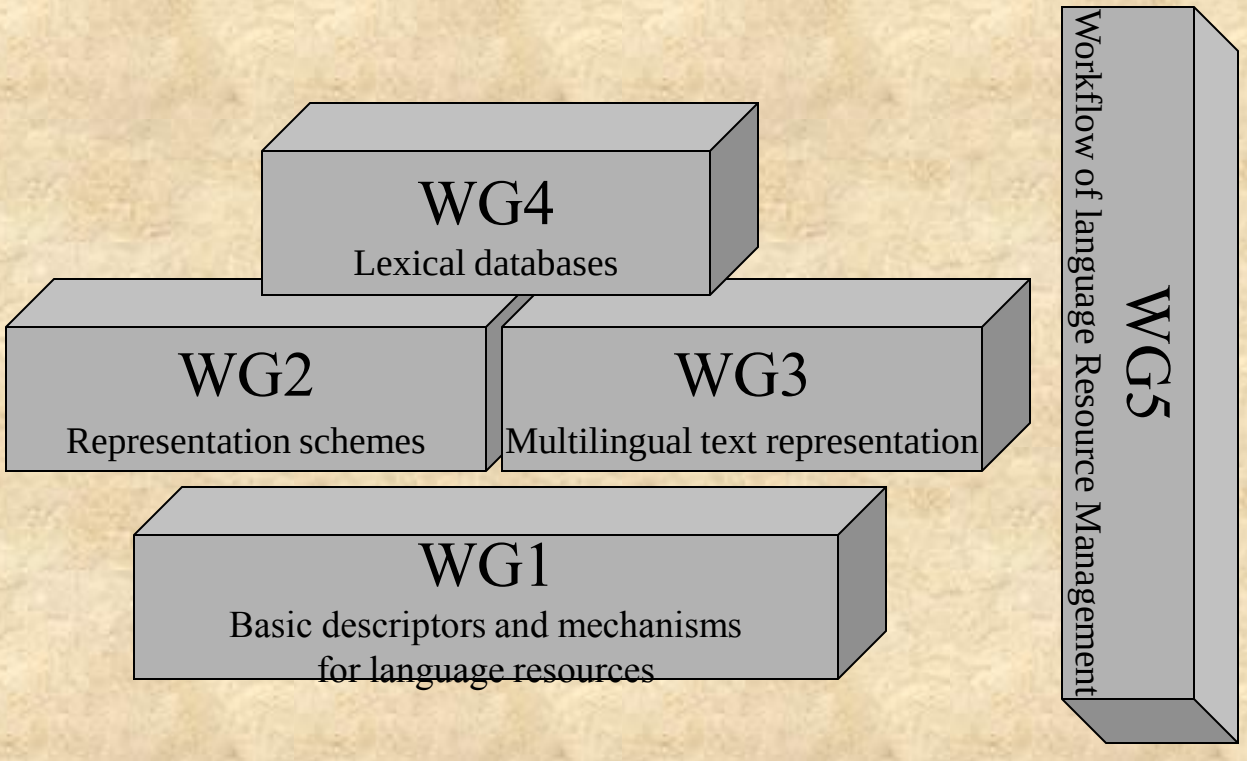
Going further

Data categories as metadata for language resources in the context of TC37 *(/SC2 + /SC3 + /SC4)

Goals of ISO TC 37/SC 4

- TC37/SC4 - Language Resource Management
  - 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.
  - Platform for designing and implementing linguistic resource formats and processes
    - Multi-layer annotation of linguistic resources
    - Exchange of information between NLP modules

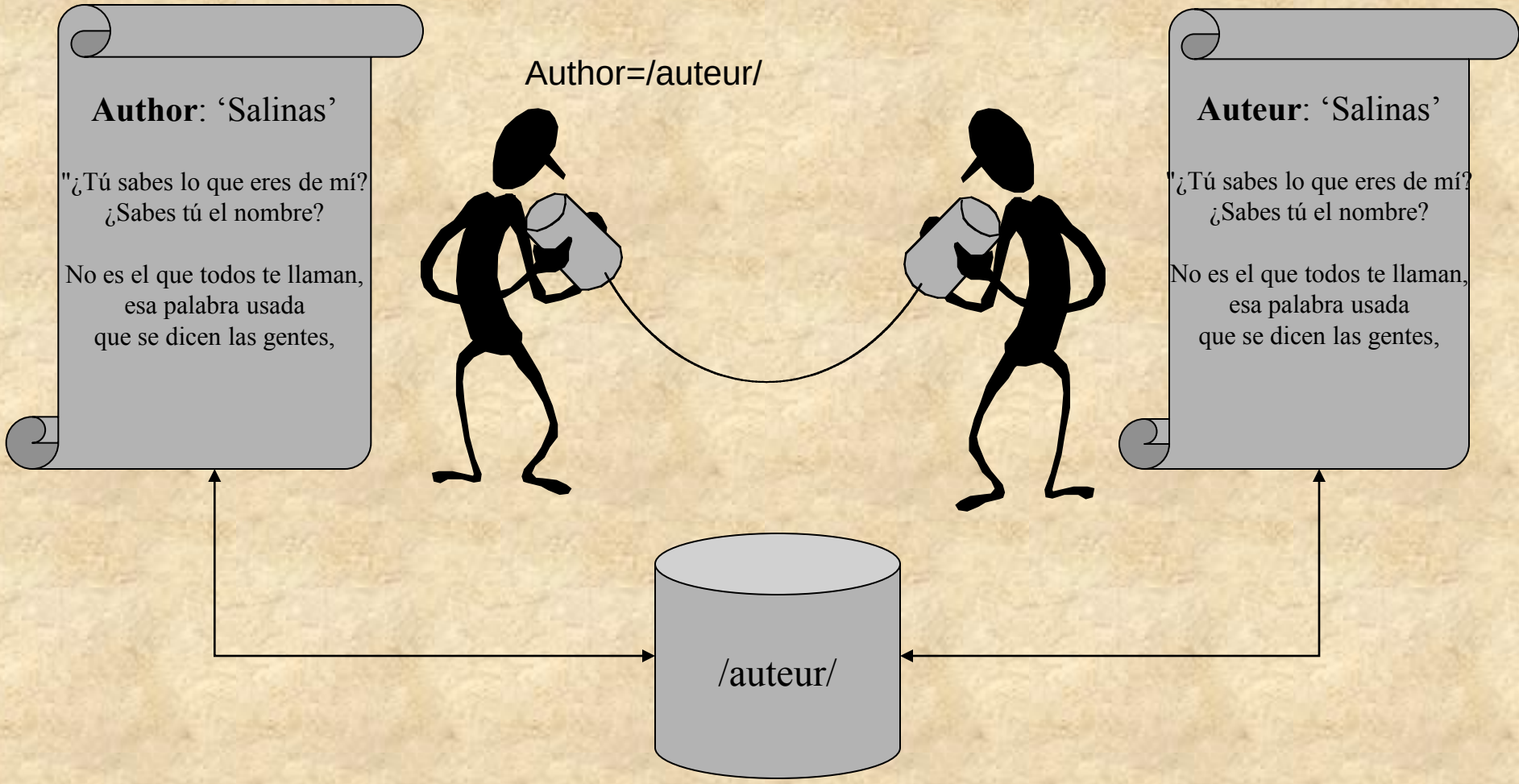
TC37/SC4 overall rationale



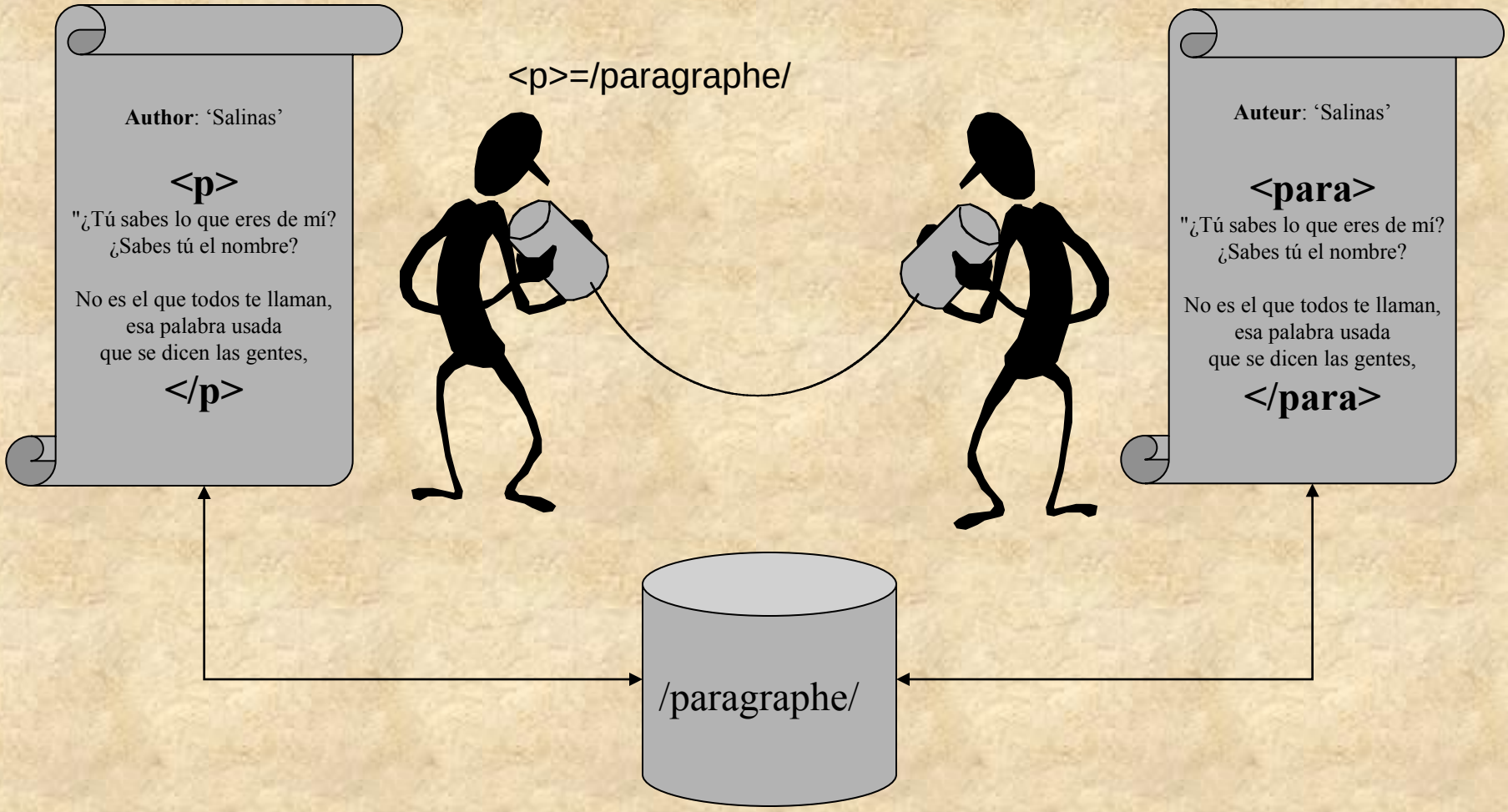
Why is metadata central?

- Problem:
  - We will never agree on one single format for one single purpose
    - Good reasons for that: various theoretical backgrounds, various levels of processing, various applicative contexts etc.
  - Standardization should provide description/mapping means between formats
    - Objective: defining interoperability principles within processing levels
      - Morpho-syntax, Syntax, Semantics, Lexica, etc.

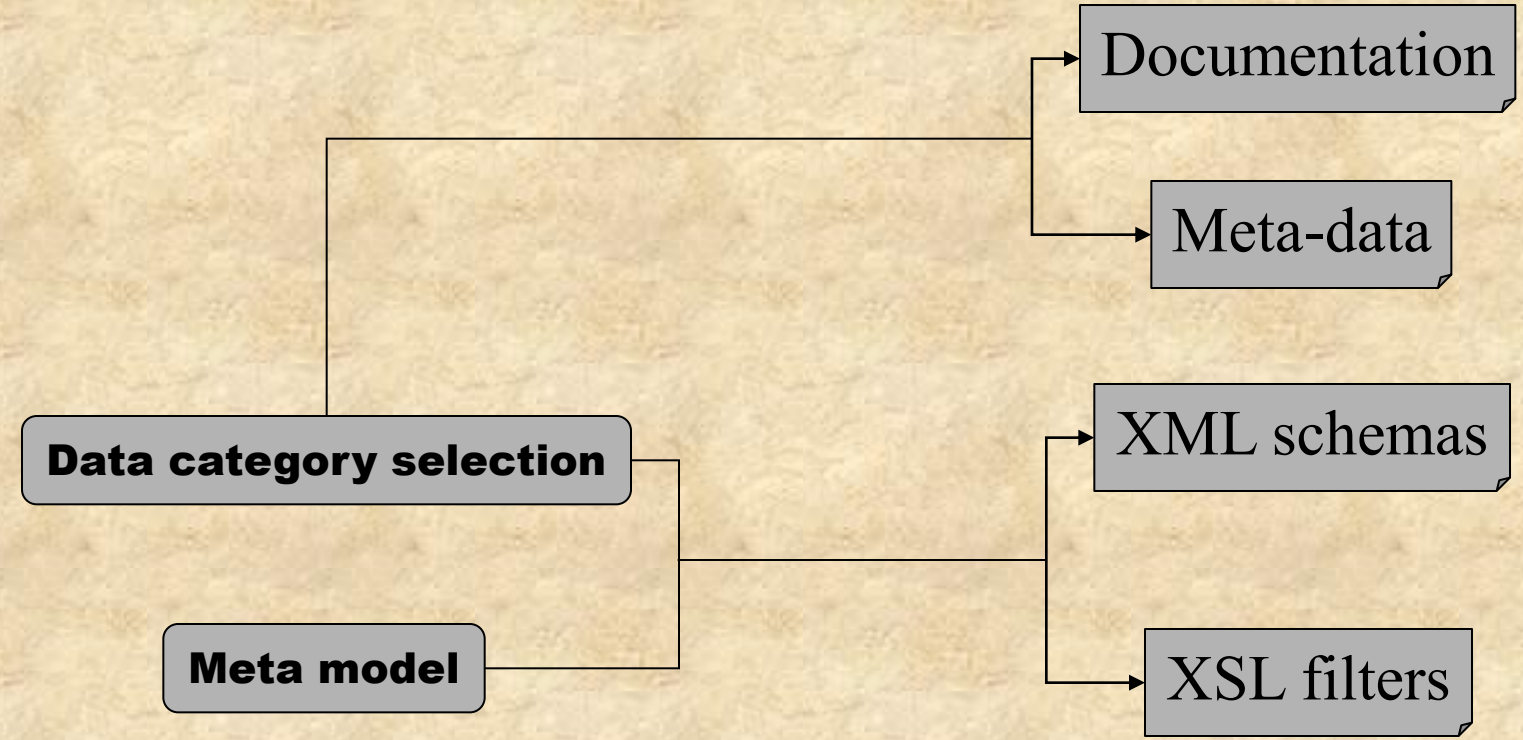
Meta data for content description



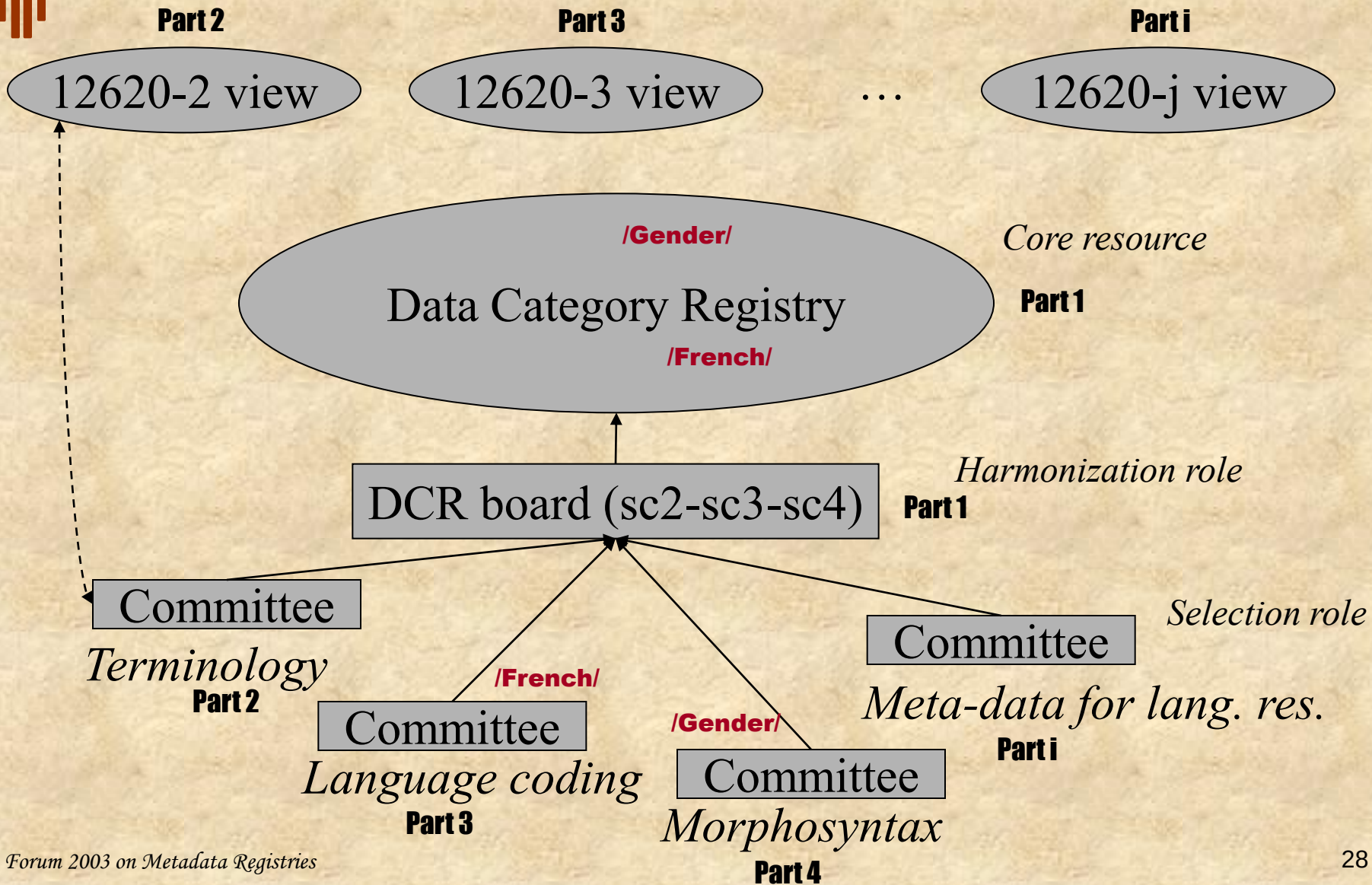
Meta data for structural description



Multiple uses of data categories



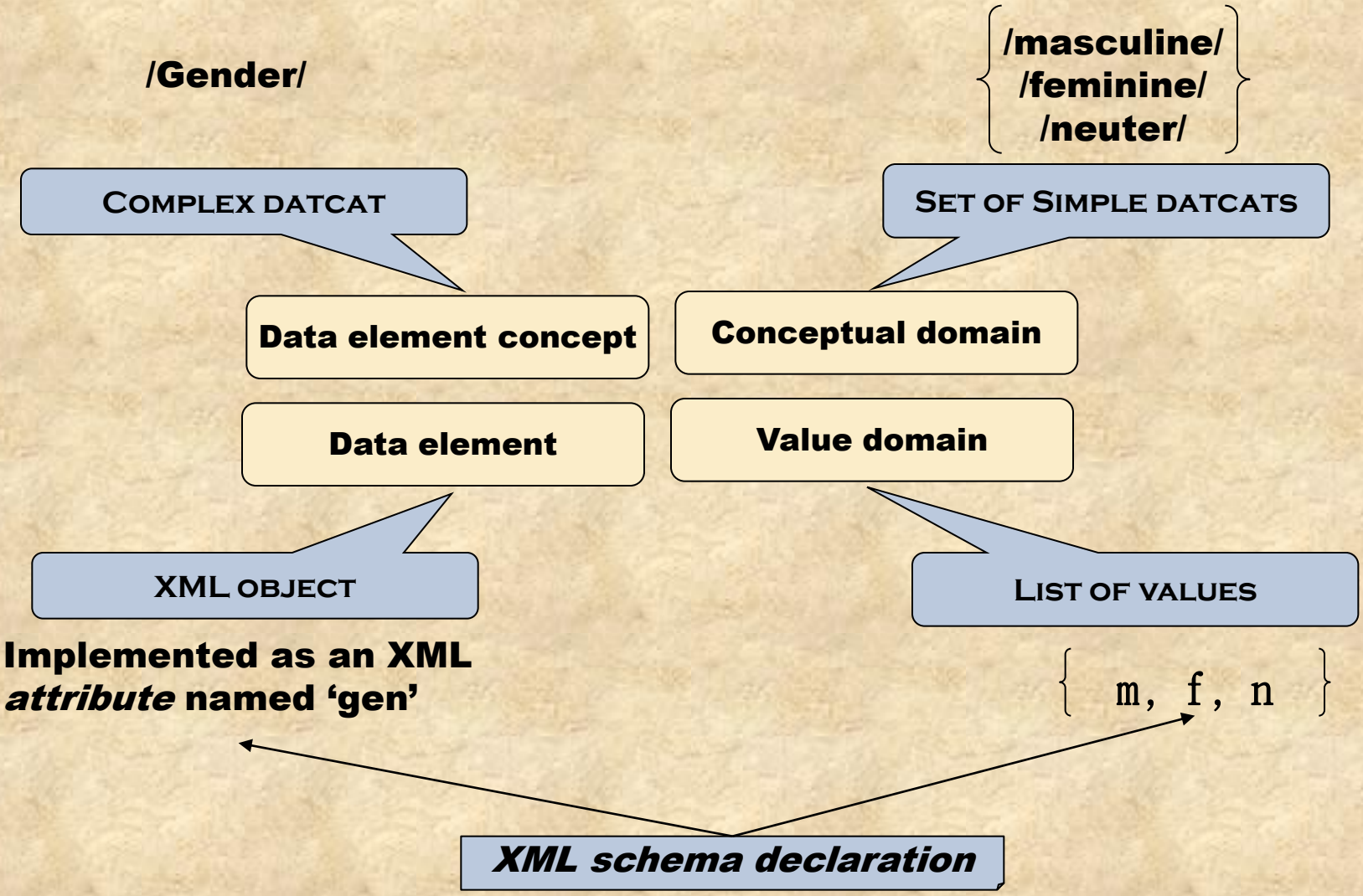
An MDR for TC37



Several issues

Understanding our relation with other initiatives

Issues - relation to ISO 11179



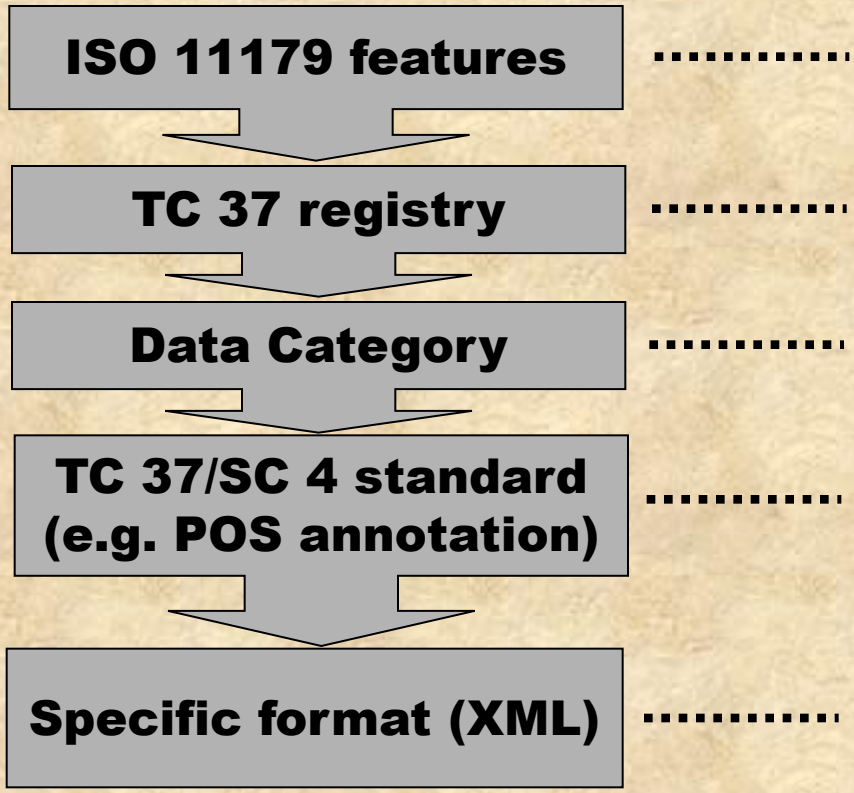
Issues

- Data categories for language resources
  - Containers and Value
    - /Gender/ → /Masculine/, /Feminine/, /Neuter/
    - Value meanings as administered items
  - Associating DatCats with views
    - Contexts?
  - Restrictions on applicability
    - /Gender/ applies to fr/en/de, but not to jp
  - Styles and vocabularies
  - Hierarchies of data categories
    - Classification system

Issues - relation to W3C

What we need to represent:

What W3C (SemWeb) Format we could use:



RDFS: to express how features combine

RDFS: specific constraints for LR

RDF: to represent Elementary entries

OWL: to relate levels in MM, properties, relations

XML schema: to control Instances of the format

Perspective

- Implementing a data category registry: a priority for TC37/SC4
  - Common background for a variety of future standards
  - Specificities related to committee activities (e.g. experts, votes)
  - Towards a real ontology of linguistic objects
- Collaboration with the ISO 11179 community is essential

For More Information

Laurent Romary

Laboratoire Loria-INRIA

Laurent.Romary@loria.fr