Mapping between the Dublin Core Abstract Model DCAM and the TMDM

TMRA 2007, Leipzig, 2007-10-12

Lutz Maicher, University of Leipzig

maicher@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Institut für Informatik

Automatische Sprachverarbeitung

Premise of the talk

- **metadata is important**
- **metadata is subject-centric**
- **DC vocabulary should be (re-)used**
- **a marriage of Topic Maps and DC is obvious**

Automatische Sprachverarbeitung

The Dublin Core Abstract Model (DCAM)

- **DC is intended to be independent from the used representation formats like RDF, HTML or Topic Maps**
  - DCAM should assure interoperability of metadata descriptions irrespective of their representation format

- **Dublin Core Abstract Model is metamodel of Dublin Core**
  - latest version April 2007 (a lot have changed!)
- **Three parts**
  - DCMI resource model
    - defines how the world is composed through DCAM glasses
  - DCMI description set model
    - specifies how information about resources are represented
  - DCMI vocabulary model
    - abstract model of the vocabulary *used* in the descriptions

UNIVERSITÄT LEIPZIG

Institut für Informatik

Automatische Sprachverarbeitung

DCMI resource model (1/2)

- **DCMI resource model specifies how the "real" world is composed through DCAM glasses**
- **the whole world is a set of *resources* (known from RDF)**
  - a resource is similar to a subject
- **each resource is composed of property / value pairs**
  - a property is a specific aspect, characteristic, attribute, or relation used to describe a resource
  - to each property a value is assigned
    - each value is by definition a resource
- **a resource becomes a "described resource" when a proxy is created**

DCMI resource model (2/2)

- **separation of *literal* values and *non-literal* values**
  - resource which is a *non-literal* value is represented by a proxy
  - resource which is a *literal* value is represented as literal

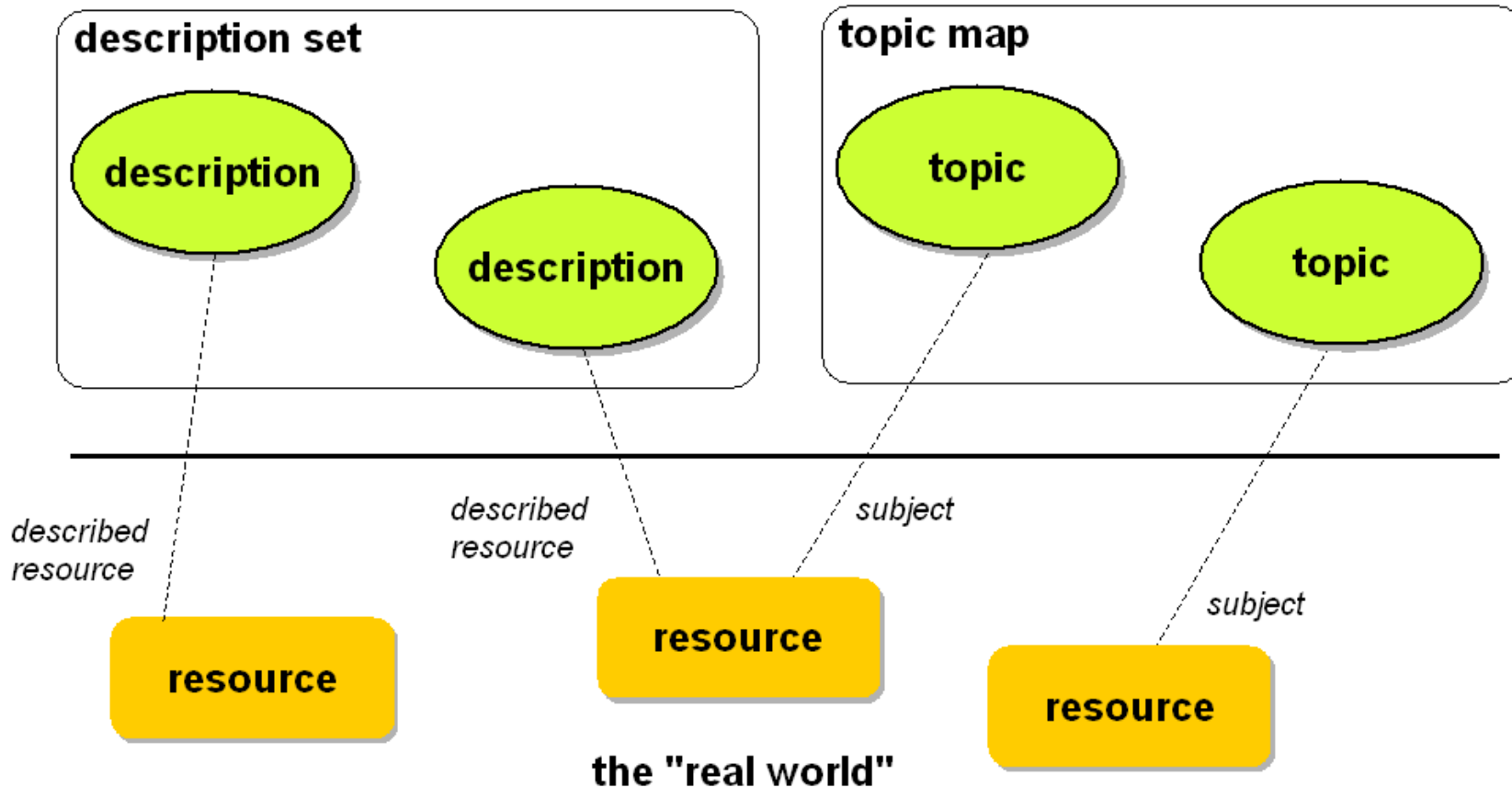*Automatische*
Sprachverarbeitung
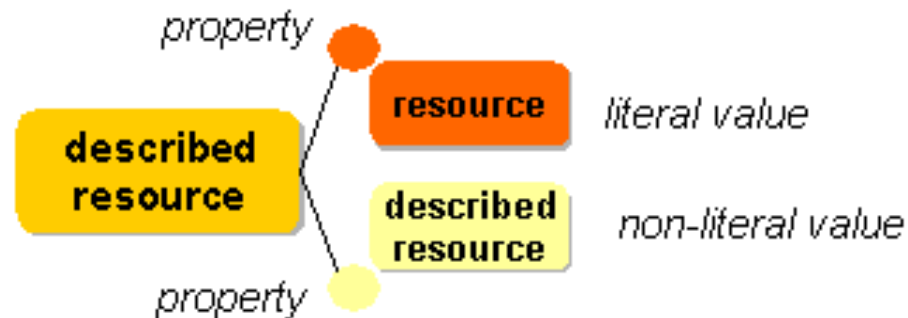
DCMI description set model (1/6)

- **DCMI description set model specifies how information about resources – which are sets of property/value pairs – will be represented**

the DCAM perspective

the TMDM perspective

description set

topic map

description

description

topic

topic

described resource

described resource

subject

subject

resource

resource

resource

the "real world"

Automatische Sprachverarbeitung

DCMI description set model (2/6)

- **The DCAM perspective**
  - *description* is the proxy of a described resource
  - *description sets* are sets of related descriptions
  - a *record* is a serialised description

- **The TMDM perspective**
  - *topic* is the proxy of a subject
  - *topic map* is a set of related subjects

DCMI description set model (3/6)

- **Descriptions are composed of:**
  - **one** described resource URI
  - a set of statements (proxies of the property/value pairs)

- **Statements as basic modelling components of the DCAM are composed of:**
  - one *property URI*
  - and *value surrogate*

- **Distinction of *Value surrogates* in**
  - literal value surrogates
  - non-literal value surrogates

*Automatische Sprachverarbeitung*

DCMI description set model (4/6)

- **non-literal value surrogate**
  - is a value URI which is a reference to the description which is the value
  - and an encoding scheme URI can be assigned
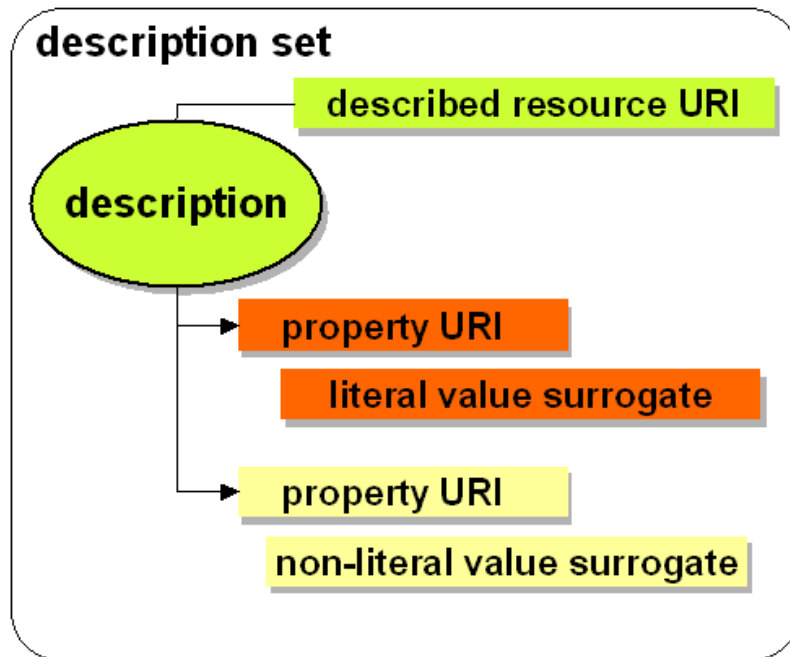   - to identify the vocabulary the used term is from

- **literal value surrogate**
  - is a value string, which is either
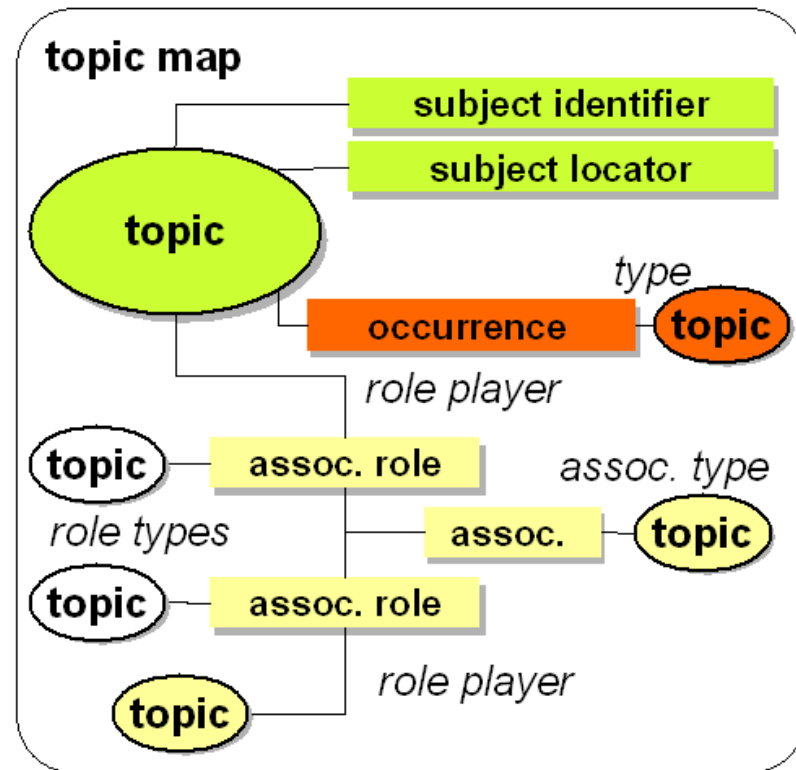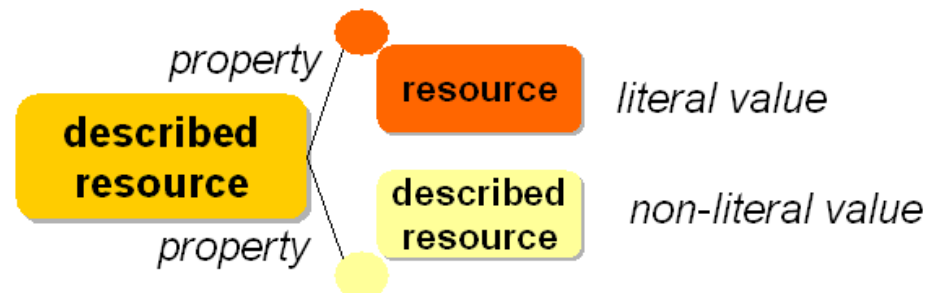   - a plain value string
    - a typed value string

UNIVERSITÄT LEIPZIG

Institut für Informatik
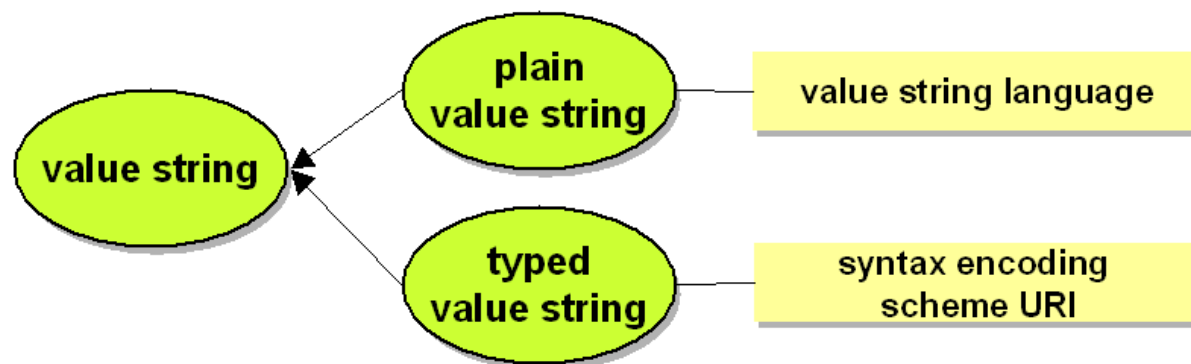
DCMI description set model (5/6)

DCMI description set model (6/6)

- **DCAM perspective**
  - described resource URI
  - statement with non-literal value surrogate
    - as proxies for non-literal value properties
  - statement with literal value surrogate
   - as proxies for literal value properties

- **TMDM perspective**
  - a set of subject identifiers / locators
  - associations
    - as proxies for non-literal value properties
  - internal occurrences
    - as proxies for literal value properties

Automatische Sprachverarbeitung

DCMI vocabulary model

- **DCMI vocabulary model is an abstract model of the vocabularies *used* in the descriptions**
  - see i.e. the 40 additional elements and refinements introduced in Steve's talk (sub-property relationships)
  - the vocabulary model is in the scope of the authoring guidelines, not in the scope of the mapping

Automatische Sprachverarbeitung

Standardising the usage of DC-vocabularies in Topic Maps

UNIVERSITÄT LEIPZIG

Institut für Informatik
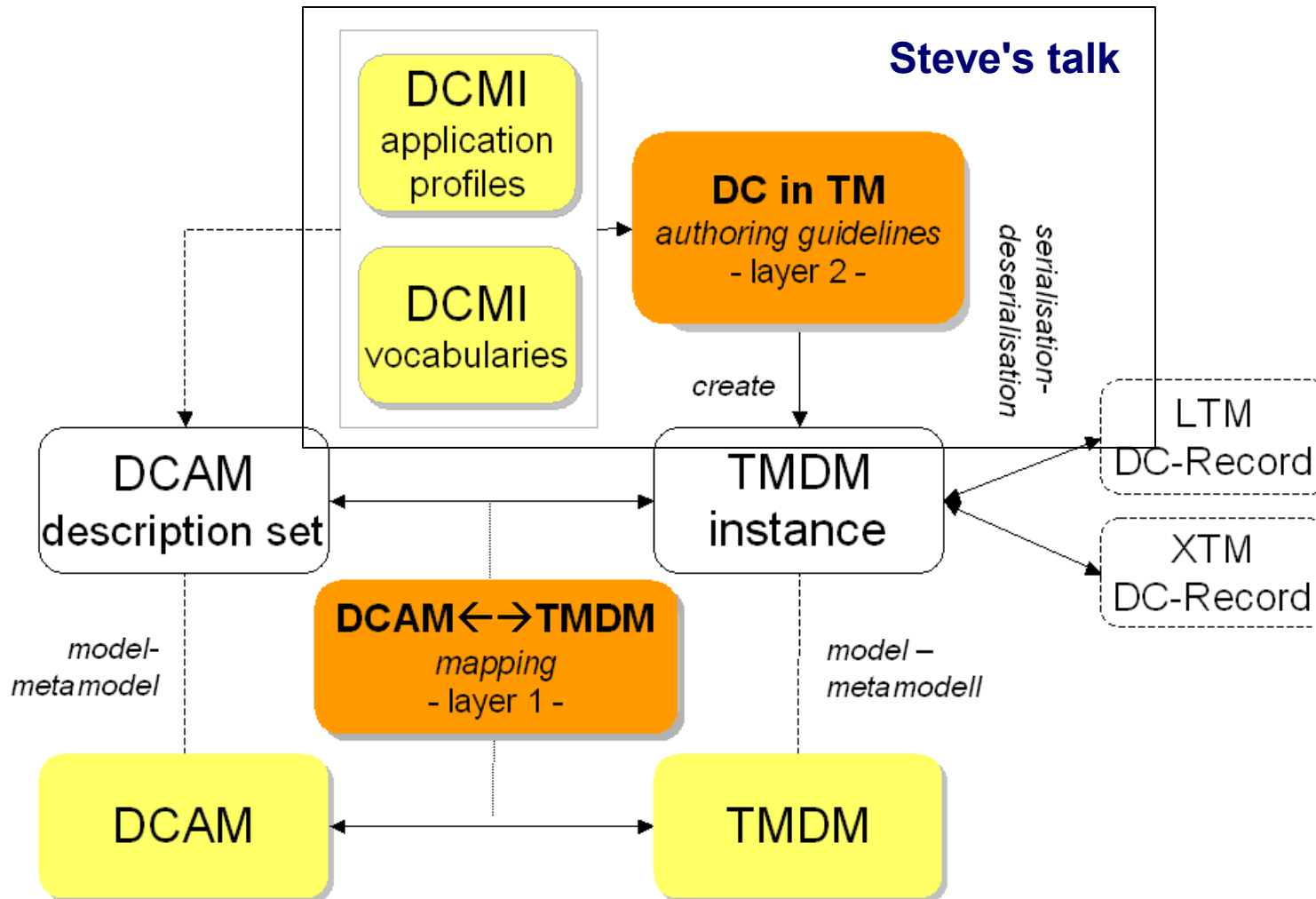
Towards authoring guidelines for DC in Topic Maps

- **For standardisation of DC/TM interoperability we forsee a two layer approach:**
  - DCAM $\leftarrow\rightarrow$ TMDM mapping as defined by this talk
    - assures general interoperability of DC metadata expressed in Topic Maps and DC metadata expressed in other representation formats

  - authoring guidlines as defined by Steve's talk
    - assures that a *authored* topic map is interoperable with DC metadata expressed in other representation format
    - assures the *mergeability* of authored topic maps (irrespective of the DCAM in the background)
    - assures that a defined set of TMQL queries can be used
    - is compliant to the DCAM $\leftarrow\rightarrow$TMDM mapping

- **DC encoding guideline for Topic Maps**
  - DCAM$\rightarrow$TMDM mapping + TMDM serialisation specifications (i.e. XTM)

DCAM and TMDM Mapping

DCAM → TMDM mapping (simplified)

- **Scope: transformation of a description set into a topic map**
- **each description D will be mapped to a topic item r**
  - described resource URI will be subject identifier or subject locator
    - as defined (recommended!) by the authoring guidelines
    - must be compliant with the TMDM semantics
  - described resource URI will be value of a typed occurrence
- **each statement S of D must be mapped**
  - if S is a non-literal value surrogate an association will be created
    - a value topic (incl. identity) must be created which represents the value
    - type of the association is topic with property URI as subject identifier
    - roles are defined terms (see Steve's talk)
  - if S is a literal value surrogate a typed occurrence must be added
    - type of occurrence is topic with property URI as subject identifier
    - value of occurrence is value string
    - add syntax encoding scheme URI as datatype of occurrence
    - add value string language as scope

DCAM → TMDM mapping (simplified)

- **Naming the resource topic → Topic Names**
  - for a better readability the value of a property which has naming characteristics can be additionally assigned as unscoped and untyped name of r
  - such topic names are only *informative*

- **Typing the resource topic**
  - for a better readability the type of a resource can be additionally assigned as type-instance relationship
  - such a type-instance relationship is only *informative*

*Automatische
Sprachverarbeitung*

TMDM→DCAM mapping (simplified)

- **Scope: transformation of topic maps (which represent metadata using the DC vocabulary) into description sets**
    - transforms only such information which is documented in the topic map using the DC vocabulary
    - all other information in a topic map is ignored
        - which uses other vocabularies
        - which uses the DC vocabulary in an incorrect way

- **this transformation might always imply a loss of information**
    - main problem: sets of subject identifiers to one described resource URI

Automatische Sprachverarbeitung

TMDM→DCAM mapping (simplified)

- **For each topic item r, which represents a described resource, a new description D must be created.**

- **A topic item is eligible:**
  – it has an occurrence item which is typed by a DC term
  – it plays a role in an association which is typed by a DC term

- **For each eligible topic, the described resource URI of D is**
  – the value of the occurrence of type dc:identifier,
  – if not available, one (randomly chosen) subject identifier or locator
  – if not available, the item identifier

- **… for more details see the paper!**

Towards authoring guidelines for DC in Topic Maps

- ***Standardised Authoring guidelines should include:***
  - *First,* it *must* be defined how a described resource and statements with literal-value and statements with non-literal values have to be represented when authoring a topic map. This standardisation must be strictly compatible to the TMDM→DCAM mapping defined here.
  - *Second,* for each term of the DC-vocabularies it *must* be decided, in which cases it should used as a property for a non-literal value and when it should be used as a property for a literal-value. **No further specifications are mandatory for any term.**
  - *Third,* guidelines for the representation of the described resources which are non-literal values *might* be defined. For example, best practice for choosing identifiers of persons, countries, dates, etc. can be defined. The more specific these specifications are, the better the mergeability of the resulting topic maps.
  - *Fourth,* it must be defined how relationships between terms of the DC-vocabulary (i.e. sub-property relationships, etc.) should be represented in Topic Maps to be compliant to the *DCMI vocabulary model*

Conclusion

- **Benefits of the defined DCAM←→ TMDM mapping**

  - assuring full interoperability between DC in Topic Maps and all other representation formats (like RDF, HTML) via the DCAM
  - providing a clear basis for the authoring guidelines
  - having DC encoding guidelines for Topic Maps
    - DCAM→TMDM + TMDM serialisation

Automatische
Sprachverarbeitung

Questions? Questions!

Lutz Maicher (maicher@informatik.uni-leipzig.de)

UNIVERSITÄT LEIPZIG

Institut für Informatik