

Dokumentieren – Validieren – Mappen - Bereinigen

Digitale Bibliothek

„Vernetzte Welten“ - Digitales Kulturerbe als öffentliches Gut

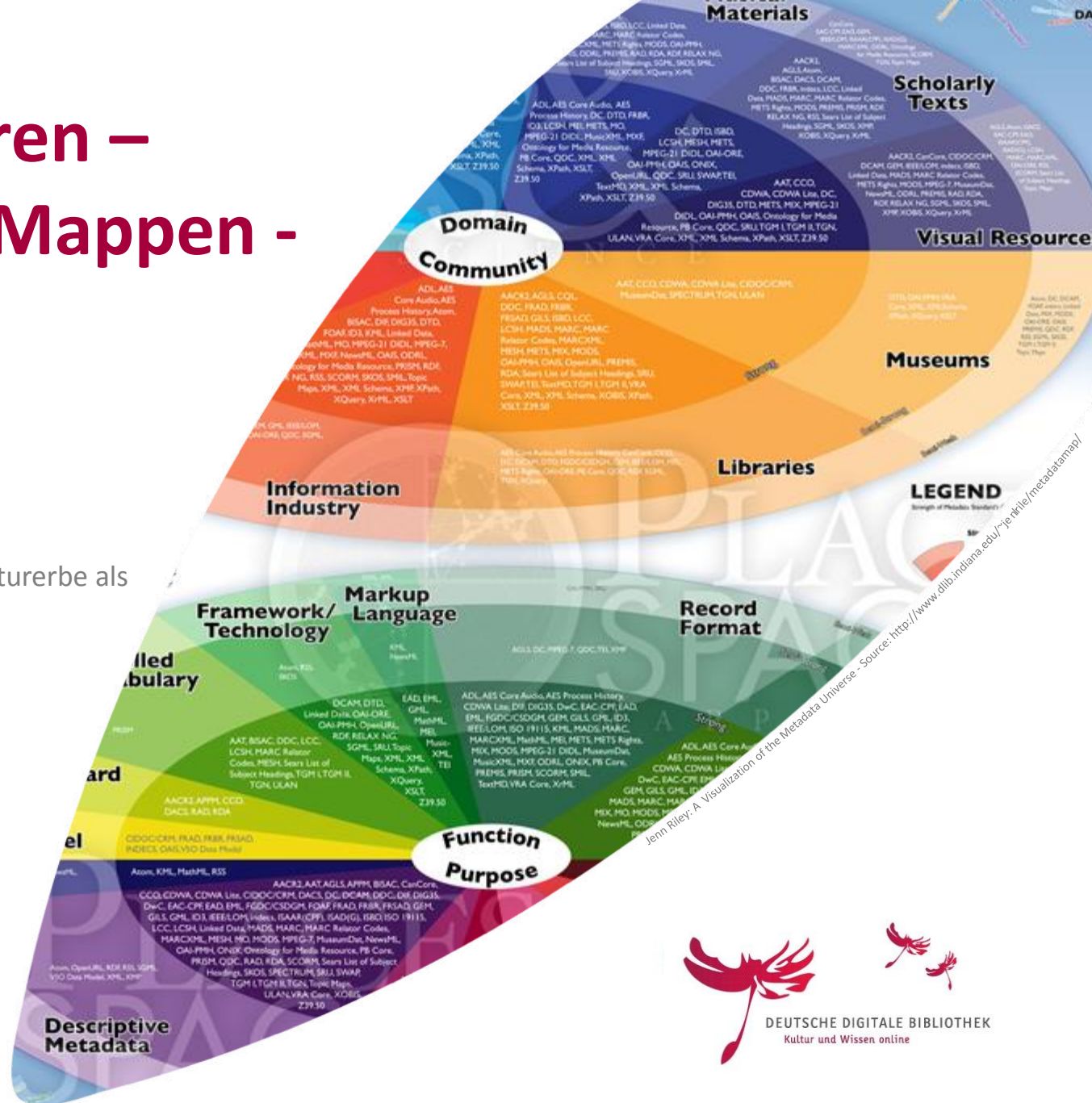
Graz, 25. Februar 2016

Stefanie Rühle

Deutsche Digitale Bibliothek
Fachstelle Bibliothek

SUB

NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN



Überblick



- Anwendungsprofile
- Validierung
- Mapping
- Bereinigen und Anreichern

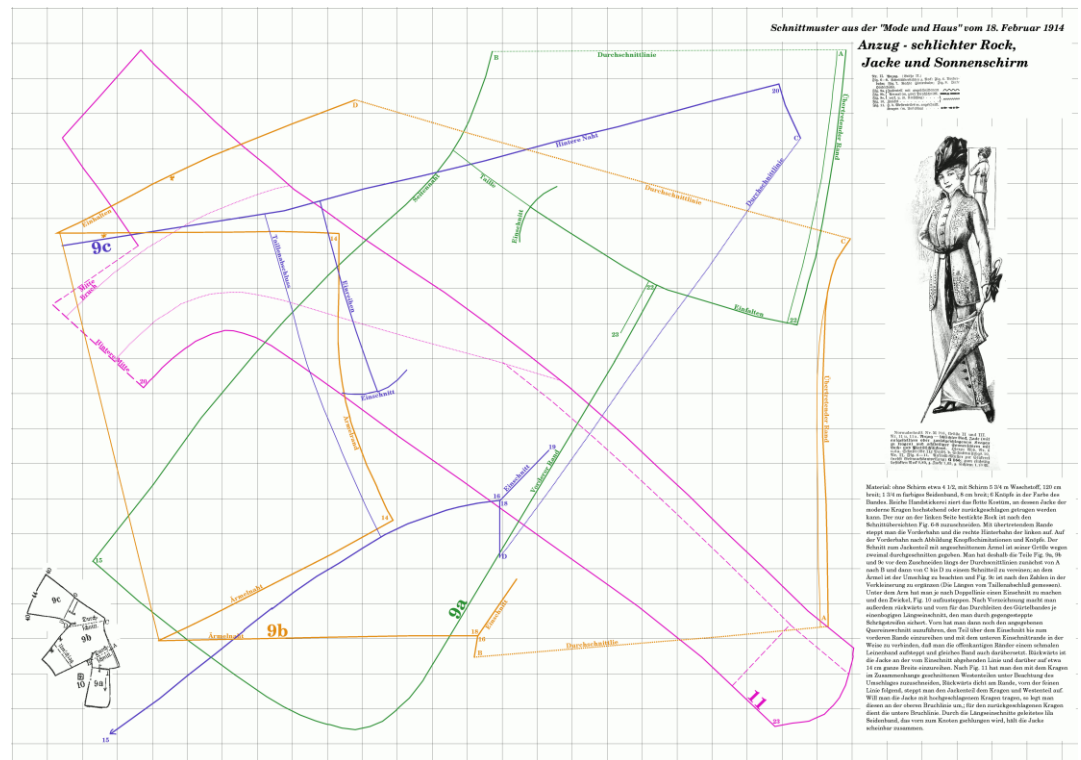
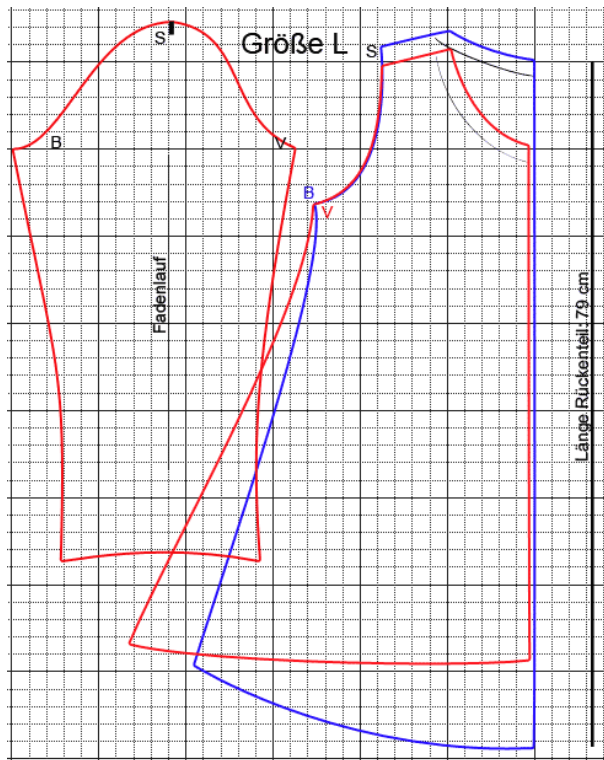
Währet den Anfängen
-> Anwendungsprofile



Anwendungsprofil / Metadatenprofil

„In the metadata community, the term application profile has been applied to describe the tailoring of standards to specific applications“

Quelle: <http://dublincore.org/documents/singapore-framework/>



Quelle: http://www.schneidern-naehen.de/files/u1/anleitungen/tunika/schnittmuster_tunika_L.gif

Quelle: <http://www.incasoftware.de/~sine/naehen/mode-und-haus/b/anzug2.gif>

Anwendungsprofil vs. LoD

- Linked open Data = Vernetzung von Daten unterschiedlichster Herkunft
- Anwendung = Software, die bestimmte Probleme lösen soll
- Anwendungsprofil = transparent machen, **was warum womit** beschrieben wird
- Grundlage ist ein **Metadatenstandard**
- Beschreibt
 - *Die Elemente, die für eine bestimmte Anwendung relevant sind*
 - *Die Entitäten, die beschrieben werden*
 - *Die Beziehungen zwischen den Entitäten*
 - *Regeln für die Verwendung der Elemente*

... Elemente, die für eine bestimmte Anwendung relevant sind

***** Subordinate Elements for 2.5.2.1 Sprache – mods:languageTerm

```
-->
<xs:element name="languageTerm" type=
<!-- -->
- <xs:complexType name="languageTermD
- <xs:simpleContent>
- <xs:extension base="stringPlusLangu
  <xs:attribute name="authorityURI
  <xs:attribute name="valueURI" ty
- <xs:attribute name="authority">
- <xs:simpleType>
- <xs:restriction base="xs:string
  <xs:enumeration value="rf
  <xs:enumeration value="is
  <xs:enumeration value="is
  <xs:enumeration value="rf
  <xs:enumeration value="rf
  <!-- -->
  </xs:restriction>
  </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="type" type="
  </xs:extension>
  </xs:simpleContent>
</xs:complexType>
- <!--
```

MODS-Definition: Contains the language(s) of the content of the resource. It may be expressed in textual or coded form. If in coded form, the source of the code is contained in the value of the authority attribute. If no authority is given, it is assumed that the content is textual.

Wiederholbar: ja
Wenn es sich um verschiedene Darstellungsformen derselben Sprache handelt.

Verpflichtungsgrad: verpflichtend
mods:language muss mindestens ein mods:languageTerm-Element, enthalten, das die Sprache in nach ISO 639-2b codierter Form enthält.

Attribute:

- **type:** Dieses Attribut gibt an, ob die Sprache in Textform oder codierter Form angegeben wird. Die Verwendung dieses Attributs ist verpflichtend. Mögliche Attributwerte sind
code für Sprachcodes und
text für Text.
- **authority:** Gibt die Codeliste an, aus der der Sprachcode entnommen wurde. Die Angabe eines Codes nach ISO 639-2b ist verpflichtend.
- **authorityURI:** URI der Codeliste.
- **valueURI:** URI des Wertes, der verwendet wird¹⁹.

Quelle: <http://www.loc.gov/standards/mods/v3/mods-3-6.xsd>

Quelle: http://dfg-viewer.de/fileadmin/groups/dfgviewer/MODS-Anwendungsprofil_2.2.pdf

... die Entitäten, die beschrieben werden

DCLib AP/DCLib Classes

Classes of the DC-Lib AP for Bibliographic Text Resources

BibliographicTextResource

Name: BibliographicTextResource

URI: <http://purl.org/dclib/terms/BibliographicTextResource> 

Definition: The intellectual creation, realization and physical embodiment of a work in text.

Type of Term: Class

DCLIB-Comment: BibliographicTextResource aggregates the properties of the FRBR entities work, expression and manifestation in one class.

Narrower Than: <http://purl.org/dc/terms/BibliographicResource> 

Narrower Than: <http://purl.org/dc/dcmitype/Text> 

Item

Name: Item

URI: <http://iflstandards.info/ns/fr/frbr/frbrer/C1004> 

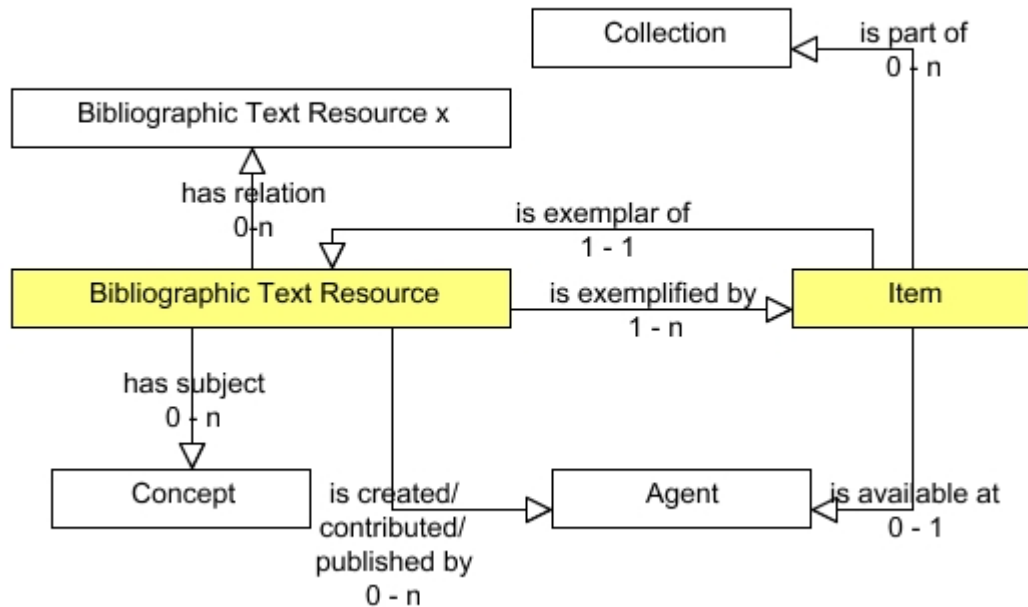
Definition: A single exemplar of a manifestation.

Type of Term: Class

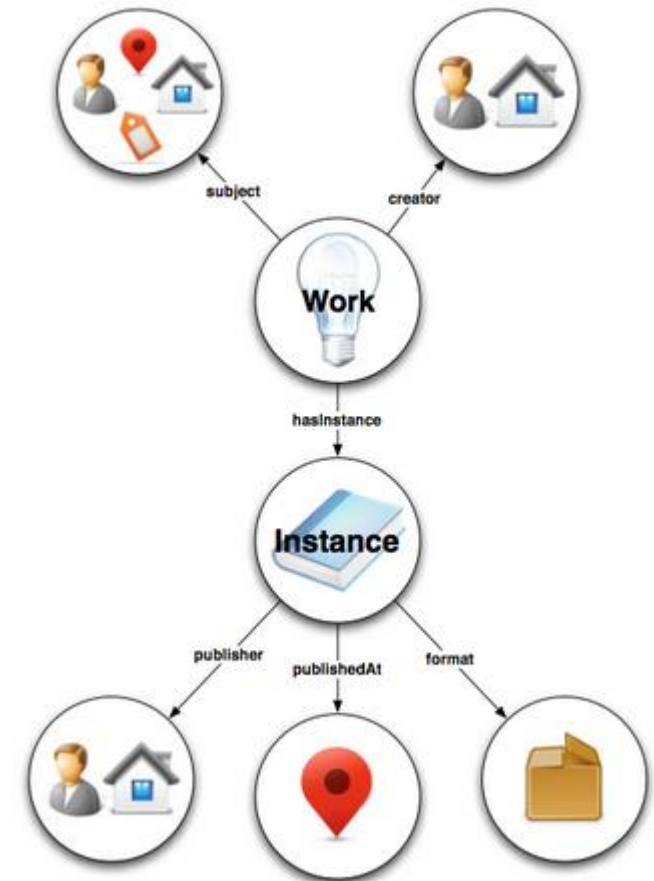
DCLIB-Comment: An item is the single copy of a Bibliographic Text Resource.

Quelle: http://wiki.dublincore.org/index.php/DCLib_AP/DCLib_Classes

... die Beziehungen zwischen den Entitäten



Quelle: http://wiki.dublincore.org/index.php/DCLib_AP



Quelle: <http://www.loc.gov/bibframe/docs/model.html>

„Rules and Constraints“

2.5.2.1 Sprache – mods:languageTerm

MODS-Definition: Contains the language(s) of the content of the resource. It may be expressed in textual or coded form. If in coded form, the source of the code is contained in the value of the authority attribute. If no authority is given, it is assumed that the content is textual.

Wiederholbar: ja

Wenn es sich um verschiedene Darstellungsformen derselben Sprache handelt.

Verpflichtungsgrad: verpflichtend

mods:language muss mindestens ein mods:languageTerm-Element, enthalten, das die Sprache in nach ISO 639-2b codierter Form enthält.

Attribute:

- **type:** Dieses Attribut gibt an, ob die Sprache in Textform oder codierter Form angegeben wird. Die Verwendung dieses Attributs ist verpflichtend. Mögliche Attributwerte sind

code für Sprachcodes und
text für Text.

- **authority:** Gibt die Codeliste an, aus der der Sprachcode entnommen wurde. Die Angabe eines Codes nach ISO 639-2b ist verpflichtend.
- **authorityURI:** URI der Codeliste.
- **valueURI:** URI des Wertes, der verwendet wird¹⁹.

Quelle: http://dfg-viewer.de/fileadmin/groups/dfgviewer/MODS-Anwendungsprofil_2.2.pdf

DCMI Application Profile

- 2003: Dublin Core Application Profile Guidelines
 - Wie sollte die Beschreibung von Elementen aussehen
 - <ftp://ftp.cenorm.be/public/cwas/e-europe/mmi-dc/cwa14855-00-2003-nov.pdf>
- 2008: Singapore Framework
 - Woraus muss ein Anwendungsprofil bestehen
 - <http://dublincore.org/documents/singapore-framework/>
- Seit Anfang 2014: RDF Application Profile Task Group
 - Wie muss eine maschinenlesbare Version eines Anwendungsprofils aussehen
 - <http://wiki.dublincore.org/index.php/RDF-Application-Profiles>

Mixing and Matching

```
<edm:ProvidedCHO
  ns3:about="http://www.deutsche-digitale-bibliothek.de/item/KNUTBNMFMBYRK6TUBGKCQB532
  xmlns:edm="http://www.europeana.eu/schemas/edm/"
  <edm:currentLocation ns3:resource="http://d-nb.info/gnd/8214-4"/>
  <edm:hasType ns3:resource="Y5PEJXNARIJHFQKEHZ6HC3SXBXUICIM4"/>
  <edm:hasMet ns3:resource="DZCTSWAALQDJUFJ32NO2RMHPACOPFXBE"/>
  <edm:hasType ns3:resource="EMCSB6Y2FKFAOILE5HTQ7R4H7Q6MN6GX"/>
  <edm:type>TEXT</edm:type>
  <dc:identifier xmlns:dc="http://purl.org/dc/elements/1.1/"
    >http://resolver.sub.uni-goettingen.de/purl?PPN729700887</dc:identifier>
  <dc:language xmlns:dc="http://purl.org/dc/elements/1.1/">ger</dc:language>
  <dc:publisher xmlns:dc="http://purl.org/dc/elements/1.1/">Salzburg:
    Duyle</dc:publisher>
  <dc:type xmlns:dc="http://purl.org/dc/elements/1.1/">Monograph</dc:type>
  <dc:type xmlns:dc="http://purl.org/dc/elements/1.1/">Ratgeber</dc:type>
  <dcterms:language ns3:resource="http://id.loc.gov/vocabulary/iso639-2/ger"
    xmlns:dcterms="http://purl.org/dc/terms/">
  <dcterms:extent xmlns:dcterms="http://purl.org/dc/terms/">[5] Bl., 118
    S.</dcterms:extent>
  <dcterms:issued xmlns:dcterms="http://purl.org/dc/terms/">1790</dcterms:issued>
  <dc:title xmlns:dc="http://purl.org/dc/elements/1.1/">F&#xfc;r Damen, und andere
    Frauenzimmer: Eine Sammlung der wichtigsten, gr&#xf6;#xdf;tentheils bisher
    geheim gehaltenen Kunstst&#xfc;rcke und Mittel, wodurch sie ihre Sch&#xf6;nheit
    erh&#xf6;hen und erhalten, verschiedene innerliche und &#xe4;u&#xdf;erliche
    Gebrechen des K&#xf6;rpers theils verh&#xfc;ten, theils heilen, und sich in
    manchen h&#xe4;uslichen Angelegenheiten leicht helfen k&#xf6;nnen</dc:title>
  <ddb:hierarchyType xmlns:ddb="http://www.deutsche-digitale-bibliothek.de/edm/"
    >type_021</ddb:hierarchyType>
  <ddb:aggregationEntity xmlns:ddb="http://www.deutsche-digitale-bibliothek.de/edm/"
    >false</ddb:aggregationEntity>
  <ddb:hierarchyPosition xmlns:ddb="http://www.deutsche-digitale-bibliothek.de/edm/"
    >DMDLOG_0000</ddb:hierarchyPosition>
</edm:ProvidedCHO>
```

Dump-Down

Properties

- dc:creator
 - dm2e:writer
 - dm2e:composer
 - dm2e:artist
- dc:identifier
 - dm2e:callNumber
- edm:hasMet
 - dm2e:sponsor
 - dm2e:principal
 - dm2e:previousOwner
 - dm2e:owner
 - dm2e:mentioned
 - dm2e:portrayed
 - dm2e:honoree
 - dm2e:holdingInstitution
 - dm2e:copyist

Property: dm2e:writer

writer

URI: <http://onto.dm2e.eu/schemas/dm2e/writer>

Domain: [edm:ProvidedCHO](#)

Range: [foaf:Person](#)

Superproperty: [dc:creator](#)

Property: dm2e:writtenAreaDimension

written area dimension

URI: <http://onto.dm2e.eu/schemas/dm2e/writtenAreaDimension>

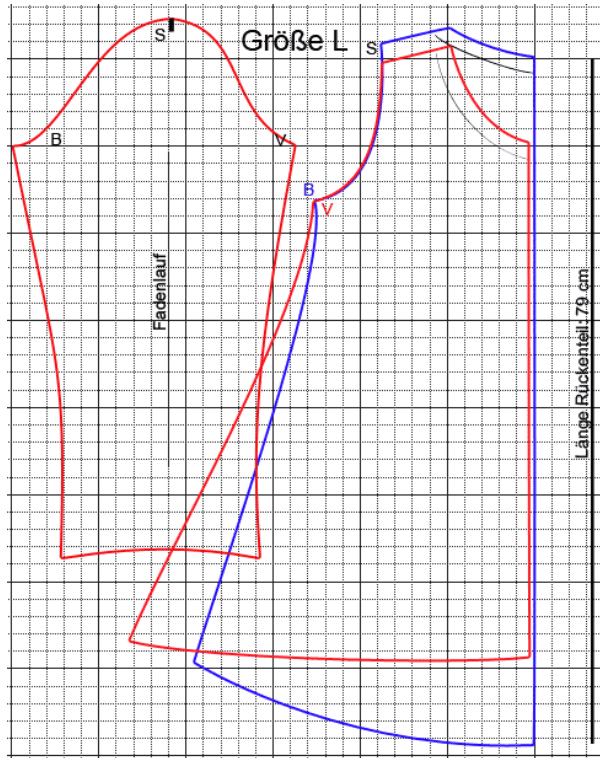
Domain: [edm:ProvidedCHO](#)

Range: [rdfs:Literal](#)

Superproperty: [dcterms:extent](#)

Quelle: <http://onto.dm2e.eu/schemas/dm2e#1.0>

Nicht die Lösung des Problems, aber



Quelle: http://www.schneidern-naehen.de/files/u1/anleitungen/tunika/schnittmuster_tunika_L.gif

- Dokumentation unterstützt
 - Datenlieferant
 - Aggregator
- Validierung -> Soll-Ist-Vergleich
- Mapping -> Transformation von einem Datenmodell in ein anderes

Validieren



- „Don't trust user input“
[https://www.owasp.org/index.php/Don't trust user input](https://www.owasp.org/index.php/Don't_trust_user_input)
 - Benutzer
 - Datenlieferung
- Daten werden auf Plausibilität geprüft
 - bei der Eingabe der Daten durch den „Nutzer“
 - bei der Übernahme von Daten aus anderen Anwendungen
- Auf der Grundlage eines
 - Schemas
 - Anwendungsprofils

Schema

- Möglichst maschinenlesbare Beschreibung der Datenstruktur
- Automatische Überprüfung der Daten hinsichtlich Konformität mit dem Schema
-> z. B. xml-Schema

```
***** Subordinate Elements for <language>

-->
<xs:element name="languageTerm" type="languageTermDefinition"/>
<!-- -->
- <xs:complexType name="languageTermDefinition">
  - <xs:simpleContent>
    - <xs:extension base="stringPlusLanguage">
      <xs:attribute name="authorityURI" type="xs:anyURI"/>
      <xs:attribute name="valueURI" type="xs:anyURI"/>
    - <xs:attribute name="authority">
      - <xs:simpleType>
        - <xs:restriction base="xs:string">
          <xs:enumeration value="rfc3066"/>
          <xs:enumeration value="iso639-2b"/>
          <xs:enumeration value="iso639-3"/>
          <xs:enumeration value="rfc4646"/>
          <xs:enumeration value="rfc5646"/>
          <!-- -->
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="type" type="codeOrText"/>
  </xs:extension>
</xs:simpleContent>
</xs:complexType>
- <!--
```

Quelle: <http://www.loc.gov/standards/mods/v3/mods-3-6.xsd>

- Prüfen der formalen Syntax -> „Läuft Haus Hans ins“

Wohlgeformtheitstest - - fehlgeschlagen. Fehler: 1

PPN630517967.mets.xml, schemata "mets.v1-7.xsd", "mods-3-3.xsd" (1 Element)

❗ - F [Xerces] The element type "mods:language" must be terminated by the matching end-tag "</mods:language>".

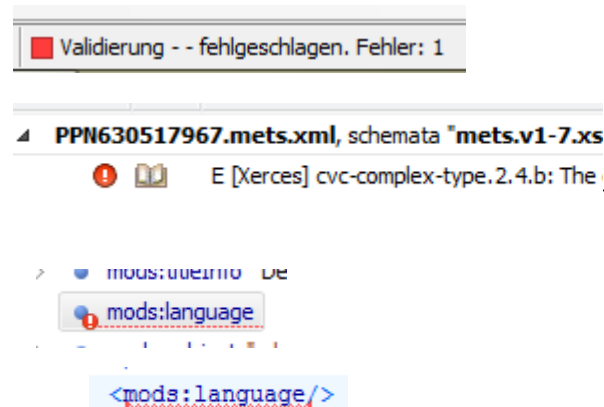
- mods:language la
 - mods:languageTerm "iso639-2b" la
- ▷ • mods:language

```
<mods:language>  
  <mods:languageTerm authority="iso639-2b" type="code">la</mods:languageTerm>  
</mods:language>
```

- Daten werden gegen das Schema geprüft

```
xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-3.xsd
http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version17/mets.v1-7.xsd">
```

- „inhaltliche“ Prüfung



```
*****
*   Top Level Element <language>                               *
*****
-->
<xs:element name="language" type="languageDefinition"/>
<!-- -->
<xs:complexType name="languageDefinition">
  <xs:sequence>
    <xs:element ref="languageTerm" maxOccurs="unbounded"/>
    <xs:element ref="scriptTerm" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="objectPart" type="xs:string"/>
  <xs:attributeGroup ref="languageAttributeGroup"/>
  <xs:attribute name="displayLabel" type="xs:string"/>
  <xs:attribute name="altRepGroup" type="xs:string"/>
  <xs:attribute name="usage" fixed="primary"/>
</xs:complexType>
<!-- -->
```

- Daten werden gegen das Anwendungsprofil geprüft
 - Manche „Fehler“ lassen sich über Schema nicht prüfen
 - Unterschiedlicher „Grad“ von Fehlern
 - Kritischer Fehler
 - Warnung
 - Hinweis
- Analyse der Feldinhalte

```
<mods:language>  
  <mods:languageTerm authority="iso639-2b" type="code">la</mods:languageTerm>  
</mods:language>
```

Schema <-> Anwendungsprofil

***** Subordinate Elements 2.5.2.1 Sprache – mods:languageTerm

```
-->
<xs:element name="languageTerm" type="languageTerm" />
<!-- -->
- <xs:complexType name="languageTerm" base="stringPlusLang" />
- <xs:simpleContent />
- <xs:extension base="stringPlusLang" />
- <xs:attribute name="authorityURI" type="anyURI" />
- <xs:attribute name="valueURI" type="anyURI" />
- <xs:attribute name="authority" type="string" />
- <xs:simpleType />
- <xs:restriction base="xs:string" />
- <xs:enumeration value="code" />
- <xs:enumeration value="text" />
- <xs:enumeration value="other" />
- <xs:enumeration value="otherCode" />
- <xs:enumeration value="otherText" />
- <!-- -->
- </xs:restriction>
- </xs:simpleType>
- </xs:attribute>
- <xs:attribute name="type" type="string" />
- </xs:extension>
- </xs:simpleContent>
- </xs:complexType>
- <!-- -->
```

MODS-Definition: Contains the language(s) of the content of the resource. It may be expressed in textual or coded form. If in coded form, the source of the code is contained in the value of the authority attribute. If no authority is given, it is assumed that the content is textual.

Wiederholbar: ja
Wenn es sich um verschiedene Darstellungsformen derselben Sprache handelt.

Verpflichtungsgrad: verpflichtend
mods:language muss mindestens ein mods:languageTerm-Element, enthalten, das die Sprache in nach ISO 639-2b codierter Form enthält.

Attribute:

- **type:** Dieses Attribut gibt an, ob die Sprache in Textform oder codierter Form angegeben wird. Die Verwendung dieses Attributs ist verpflichtend. Mögliche Attributwerte sind *code* für Sprachcodes und *text* für Text.
- **authority:** Gibt die Codeliste an, aus der der Sprachcode entnommen wurde. Die Angabe eines Codes nach ISO 639-2b ist verpflichtend.
- **authorityURI:** URI der Codeliste.
- **valueURI:** URI des Wertes, der verwendet wird¹⁹.

- Validiert Struktur und Inhalt von XML-Dokumenten
- XML-Dokument mit verschiedenen Validierungsregeln
- Ergänzung zu XML-Schema
- Ermöglicht Regeln zu definieren, die mit XML-Schema nicht möglich sind

```
<p class="pathContainer">
  <ul>
    <li>
      <p class="path">
        <code>
          <font size="-2" color="#808080">mods:language/mods:languageTerm[@authority='iso639-2b']
            [not(document('iso639_2.rdf')//madsrdf:code-normalize-space(current()))]</font>
        </code>
      </p>
      <p class="message">
        'mods:languageTerm' hat einen falschen Sprachcode:
        'la'<br/> - Vorkommnisse: 2<br/> - Beispiel: identifier 'PPN630517967' , mets:dmdSec[@ID='DMDLOG_0000']</p>
      </li>
    </ul>
  </p>
```

Mapping



- Ausgangslage: Datenbestände aus unabhängigen, heterogenen Systemen in ein gemeinsames Datenformat zusammenführen
 - Übersetzung von Elementen der einen Datenstruktur in die einer anderen
 - Transformation von „Feldinhalten“ aus einer Datenstruktur in eine andere
 - Auf der Grundlage des kleinsten gemeinsamen Nenners
- Probleme:
 - Unterschiedliche Zeichencodierung (Latin 1, UTF 8, ...)
 - Unterschiedliche Codes für z. B. Sprachen, Länder, ...
 - Textelement in einer Sprache in eine andere übersetzen (z. B. Bild -> Image)
 - Redundanzen beseitigen
 - Felder zerlegen (z. B. Titelinformationen)
 - Mehrere Felder zusammenfügen
 - Einzelne Datensätze in mehrere Datensätze zerlegen (z. B. Mehrbändige Werke)
 - Mehrere Datensätze zu einem Datensatz zusammenfassen
 - Verknüpfungsstrukturen (Normdaten, Teil-Ganzes-Beziehungen)
 - ...

Beispiel PICA+ -> DCTERMS

PICA+	Definition PICA	DCTERMS	Definition DCMES
021A\$a	Hauptsachtitel	dcterms:title	A name given to the resource.
021A\$d	Zusätze zum Hauptsachtitel		
021A\$f	Parallelsachtitel	dcterms:alternative	An alternative name for the resource.
028A\$a	Verfasser Familienname	dcterms:creator	An entity primarily responsible for making the resource.
028A\$d	Verfasser Vorname		
029A\$a	Name der Körperschaft	dcterms:contributor	An entity responsible for making contributions to the resource.
029A\$b	Abteilung der Körperschaft		
033A\$p	Ort		
033A\$n	Verlag	dcterms:publisher	An entity responsible for making the resource available.

Mapping

- Vom Besonderen zum Allgemeinen
 - rdavocab:variantTitleForTheWork -> dc:title
- Vom Allgemeinen zum Allgemeinen
 - rdavocab:title -> dc:title
- Vom Allgemeinen zum Besonderen
 - dc:contributor -> mods:name@type="personal" oder mods:name@type="corporate"
- Vom Besonderen zum Besonderen
 - schema:startDate -> dcterms:issued oder dcterms:modified oder dcterms:created
- Mapping ist aufwendig und führt häufig zu Informationsverlust

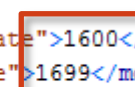
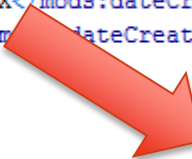
Bereinigen und Anreichern



Daten bereinigen

- Bereinigen
 - Maschinenlesbarkeit der Daten verbessern

```
<mods:originInfo>
  <mods:dateCreated encoding="w3cdtf" keyDate="yes">16XX</mods:dateCreated>
  <mods:dateCreated qualifier="approximate">[17. Jh.]</mods:dateCreated>
</mods:originInfo>
<mods:originInfo>
  <mods:dateCreated encoding="iso8601" point="start" qualifier="approximate">1600</mods:dateCreated>
  <mods:dateCreated encoding="iso8601" point="end" qualifier="approximate">1699</mods:dateCreated>
  <mods:dateCreated qualifier="approximate">[17. Jh.]</mods:dateCreated>
</mods:originInfo>
```



- Vereinheitlichung der Daten

```
<mods:language>
  <mods:languageTerm authority="iso639-2b" type="code">la</mods:languageTerm>
</mods:language>
```



```
<mods:language>
  <mods:languageTerm authority="iso639-2b" type="code">lat</mods:languageTerm>
</mods:language>
```

- „Neue“ Informationen in die Daten übernehmen
- Beispiel:
 - Übernahme von Identifiern in die Daten (z. B. der Name des Datenlieferanten wird durch ISIL-URI ersetzt um ihn auch in anderen Zusammenhängen eindeutig identifizieren zu können)
 - Übernahme von Informationen aus anderen Datensätzen (z. B. Erscheinungsjahr des Zeitschriftenbandes wird in Aufsatzbeschreibung übernommen)
 - Spezifizierung, wenn die Wiederholung eines Elements einen anderen Sachverhalt ausdrückt (z. B. zwei Datumsangaben, von denen sich eine auf das Original, die andere auf das Digitalisat bezieht, werden als solche gekennzeichnet)
 - ...
- Automatisierte Datenanreicherung ist fehleranfällig, manuelle Datenanreicherung teuer

Fazit



- Anwendungsprofil = Dokumentation
 - Unterstützt Datenlieferanten und Aggregatoren
 - Macht deutlich, wie die Daten aussehen sollen -> Mapping
 - Maschinenlesbares AP erleichtert Validierung
- Schemata ermöglichen die automatisierte Prüfung großer Datenmengen auf Konformität mit Standards und Aps
- Mapping und Datenbereinigung sind zeitaufwendig, fehleranfällig und haben nicht selten Informationsverlust zur Folge

Vielen Dank

Stefanie Rühle

sruehle(a)sub.uni-goettingen.de

DDB Fachstelle Bibliothek

bibliothek(a)deutsche-digitale-bibliothek.de

<http://www.deutsche-digitale-bibliothek.de>

<https://pro.deutsche-digitale-bibliothek.de>

www.facebook.com/ddbkultur

www.twitter.com/ddbkultur