berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Metadaten – Nutzen und Nutzung

Axel Herold

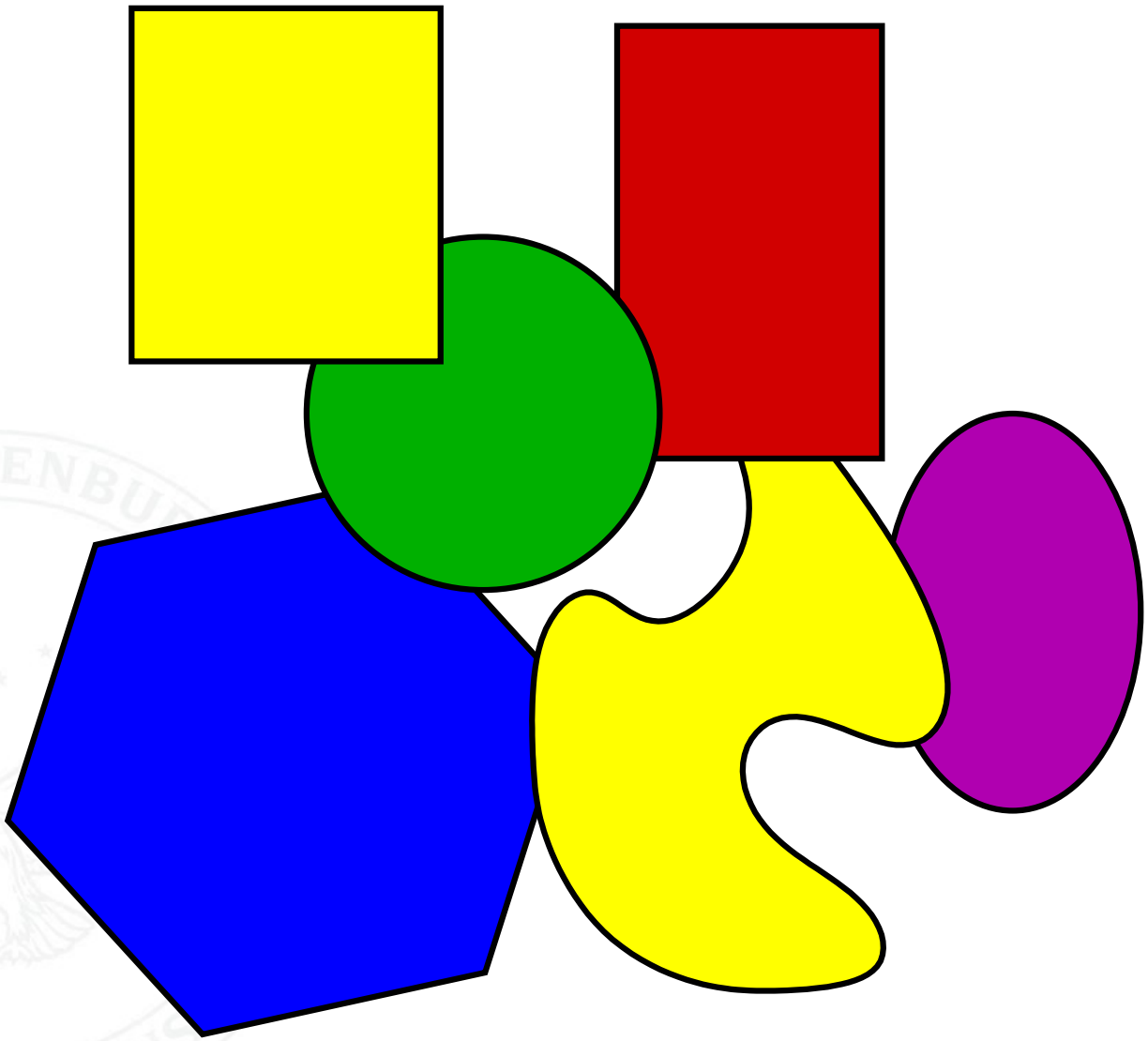
Berlin-Brandenburgische Akademie der Wissenschaften

18. Februar 2013, DTA/CLARIN-D-Konferenz

Übersicht

1. Was sind eigentlich Metadaten?
2. Wozu werden Metadaten benötigt?
3. Metadaten in CLARIN-D
   - Metadaten-Infrastruktur
   - Metadaten-Erstellung
   - Anwendungen

Metadaten, I



Metadaten, II

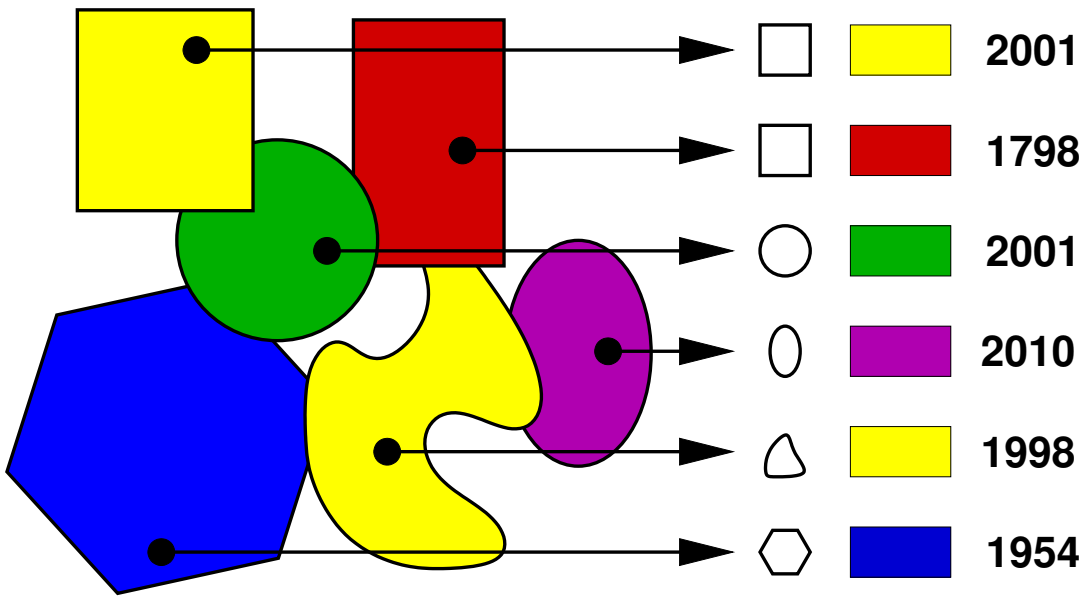
berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

(formalisierte) Beschreibungen von . . . „Dingen“

formalisiert – vereinbarte Abstraktion wegen automatischer Verarbeitbarkeit, Klassenbildung

Beschreibung – Aufzählung von relevanten **Eigenschaften**

Ding – hier: linguistische Daten, Werkzeuge

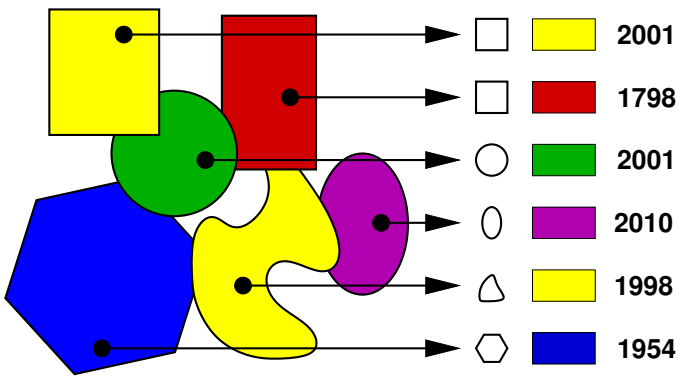


Metadaten, III

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Metadaten als Aufzählungen von Eigenschaften

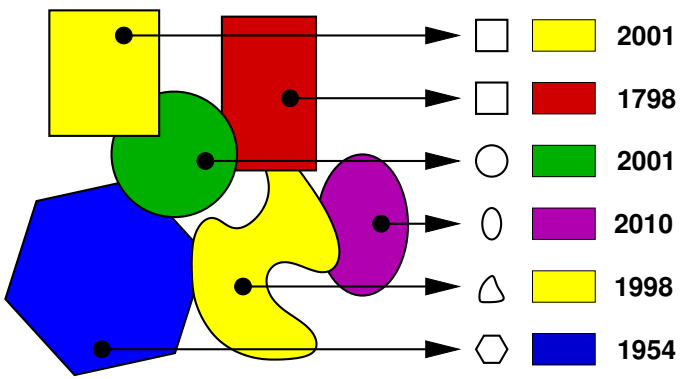
- Was ist für mich wichtig?
- Was ist für andere wichtig?
  - Forscher:
    wiederfinden, zitieren, Rahmenbedingungen, ...
  - Forschungsfragen:
    angemessene, „passende“ Daten, ...
  - Methoden/Verfahren:
    Kompatibilität, Objektrepräsentation, ...



Anwendungen

Beispiele

- Bibliotheks-, Archivkataloge, ...
- Archivierung (elektronische Repositorien)
- Versionierung
- Ressourcen-, Werkzeugsuche
  (siehe Virtual Language Observatory,
  `http://www.clarin.eu/vlo`)
- Bedingungen für Reihenschaltung von Werkzeugen
  (siehe WebLicht-Beitrag)



Modellierung von Metadaten, I

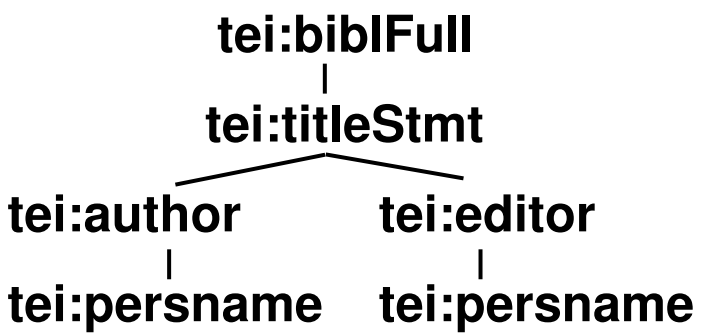
(formalisierte) Beschreibungen von . . . „Dingen“

vektoriell: Tupel fixer Feld- und Datentypen, explizite Semantik
(DCMI element set, DCMI metadata terms)

baumartig: hierarchische Anordnung von Feldern,
explizite und implizite Semantik
(TEI-Header)

modular: Mengen von Tupeln und/oder Bäumen, explizite Semantik
(CMDI+ISOcat(+RelCat))

dc:title	**...**
dc:identifier	**...**
dc:language	**...**
...	**...**



Modellierung von Metadaten, II

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

CMDI: modulare Metadaten

- Component MetaData Infrastructure, `http://www.clarin.eu/cmdi`
- explizite Semantik via ISOcat (`http://www.isocat.org/`)
- Komponentenregistratur (`http://catalog.clarin.eu/ds/ComponentRegistry/`) bedienbar per Webfrontend und als Webservice
- verschiedene Editoren, Konverter (Arbil, proforma, . . .)

Modellierung von Metadaten, III

CMDI kurz und knapp: Meta-Metadatenmodell

- Konstruktionssystem für Metadatenformate (**Profile**) aus einzelnen wiederverwendbaren **Komponenten**
  $\longrightarrow$ „Baukastenprinzip“
- kann alle existierenden Formate abbilden
- kann beliebige „Dinge“ beschreiben
- beliebige Granularität der Beschreibung
- unabhängige Komponenten erlauben konkurrierende Klassifikationen
- Serialisierung der Metadateninstanzen in XML
- Schemata für Instanzen dynamisch erzeugt

CMDI im Detail I

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Integration existierender Formate

Name: teiHeader
Group Name: CLARIN–D: DTA–Basisformat
Description: the version of the teiHeader
that is used by the DTA project, see
http://www.deutschestextarchiv.de/doku/basisformat_header

Component: fileDesc

Number of occurrences: 1–1

Component: encodingDesc

Number of occurrences: 1–1

Component: profileDesc

Number of occurrences: 1–1

CMDI im Detail II

Name: fileDesc
Group Name: CLARIN–D: DTA–Basisformat
Description: metadata for the electronic edition of a text

Component: titleStmt
Number of occurrences: 1–1

Component: editionStmt
Number of occurrences: 1–1

Component: extent
Number of occurrences: 1–1

Component: publicationStmt
Number of occurrences: 1–1

Component: sourceDesc
Number of occurrences: 1–1

CMDI im Detail III

Name:extent
Group Name: CLARIN–D: DTA–Basisformat
Description: the size of a resource with respect
to a specified unit of measurement

Element: measure string
Number of occurrences: 1–unbounded
AttributeList:
type: {images, tokens, types, characters, pages}

CMDI im Detail IV

XML-Serialisierung

```
<?xml version="1.0" encoding="UTF-8"?>
<CMD CMDVersion="1.1" xmlns="http://www.clarin.eu/
    cmd/" xmlns:xsi="http://www.w3.org/2001/
    XMLSchema-instance">
  <Header>
    <MdCreator>Deutsches Textarchiv</MdCreator>
    <MdCreationDate>2012-12-06</MdCreationDate>
    <MdSelfLink>http://www.deutschestextarchiv.de/
        api/cmdi/schwab_sagen03_1840</MdSelfLink>
    <MdProfile>clarin.eu:cr1:p_1345180279115</
        MdProfile>
    <MdCollectionDisplayName/>
  </Header>
  <Resources>...</Resources>
  <Components>...</Components>
</CMD>
```

CMDI im Detail V

```
<CMD>
  <Header>...</Header>
  <Resources>
    <ResourceProxyList>
      <ResourceProxy id="xml">
        <ResourceType mimetype="application/xml">
            Resource</ResourceType>
        <ResourceRef>http://www.deutschestextarchiv
            .de/book/download_xml/
            schwab_sagen03_1840</ResourceRef>
      </ResourceProxy>
      <ResourceProxy>...</ResourceProxy>
    </ResourceProxyList>
    <JournalFileProxyList></JournalFileProxyList>
    <ResourceRelationList></ResourceRelationList>
    <IsPartOfList></IsPartOfList>
  </Resources>
  <Components></Components>
</CMD>
```

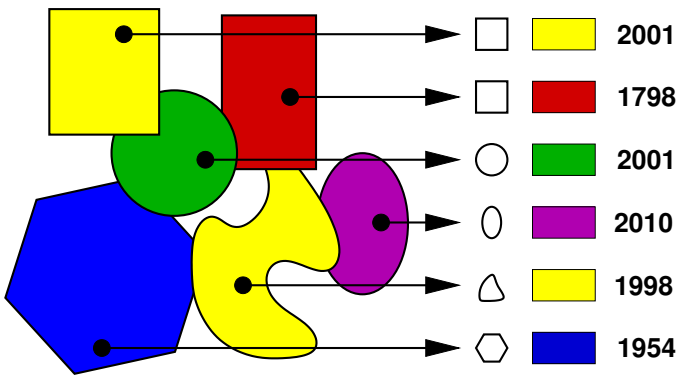
CMDI im Detail VI

```
<CMD>
  <Header>...</Header>
  <Resources>...</Resources>
  <Components>
    <teiHeader>
      <fileDesc>
        <titleStmt>
          <title type="main">Die schönsten Sagen
              des klassischen Alterthums</title>
          <title type="volume" n="3">Dritter Theil<
              /title>
          <author>
            <persName>
              <surname>Schwab</surname>
              <forename>Gustav</forename>
              <idno><idno type="PND">http://d-nb.
                  info/gnd/118762745</idno></idno>
            </persName>
          </author> ...
```

CMDI im Detail VII

Entwurfsprinzipien

1. Profile wiederverwenden
2. Komponenten wiederverwenden
3. existierende Komponenten modifizieren
4. eigene Komponente entwickeln
   (dabei auf Wiederverwendbarkeit achten)



CMDI im Detail VIII

Beispiel: Wörterbücher

Name: LexicalResource
Group Name: CLARIN–D: Lexical resource
Description: a profile for describing a lexical resource
Concept Link: http://www.isocat.org/datcat/DC–3296

Component: ExternalProperties
Number of occurrences: 1–1

Component: InternalProperties
Number of occurrences: 1–1

Component: cmdi–description
Number of occurrences: 0–1

CMDI im Detail IX

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Name: ExternalProperties
Group Name: CLARIN–D: Lexical resource
Description: properties of entities that are related to the resource
but external to it, i.e. not part of the resource proper

Component: cmdi–generalinfo

Number of occurrences: 1–1

Component: cmdi–creators

Number of occurrences: 1–1

Component: cmdi–project

Number of occurrences: 0–unbounded

Component: cmdi–access

Number of occurrences: 0–unbounded

...

CMDI im Detail X

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Name: InternalProperties
Group Name: CLARIN–D: Lexical resource
Description: properties of the resource proper

Component: ResourceType
Number of occurrences: 1–1

Component: cmdi–annotationtypes
Number of occurrences: 0–1

Component: cmdi–subjectlanguages
Number of occurrences: 0–1

Component: cmdi–modality
Number of occurrences: 0–1

Component: cmdi–description
Number of occurrences: 0–1

CMDI im Detail XI

Beispiel: Werkzeug

Name: Tool
Group Name:
Description: Description of a tool

Element: toolType string
Concept Link: http://www.isocat.org/datacat/DC–3810
Number of occurrences: 1–unbounded

Element: applicationType string
Concept Link: http://www.isocat.org/datacat/DC–3786
Number of occurrences: 1–1

Component: cmdi–generalinfo
Number of occurrences: 1–1

Component: cmdi–creators
Number of occurrences: 1–1

...

Component: toolInput
Number of occurrences: 0–1

...

CMDI in CLARIN-D

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Repositorien

CMDI-Metadaten (über OAI-PMH) sind eine Mindestanforderung

Virtual Language Observatory

- periodische Abfrage von Repositorien
- Aggregator und Suchmaschine (*facetted search*)
- Unifikation von Kategorien über ISOcat
- `http://www.clarin.eu/vlo/`

WebLicht

- Ermitteln möglicher Reihenschaltungen von Werkzeugen
- `toolOutput` von $W_1 \equiv$ `toolInput` von W_2?
- `https://weblicht.sfs.uni-tuebingen.de/`

Vielen Dank!