Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

DARIAH-DE
Digital Research Infrastructure
for the Arts and Humanities

Referat 02:
Markup und XML

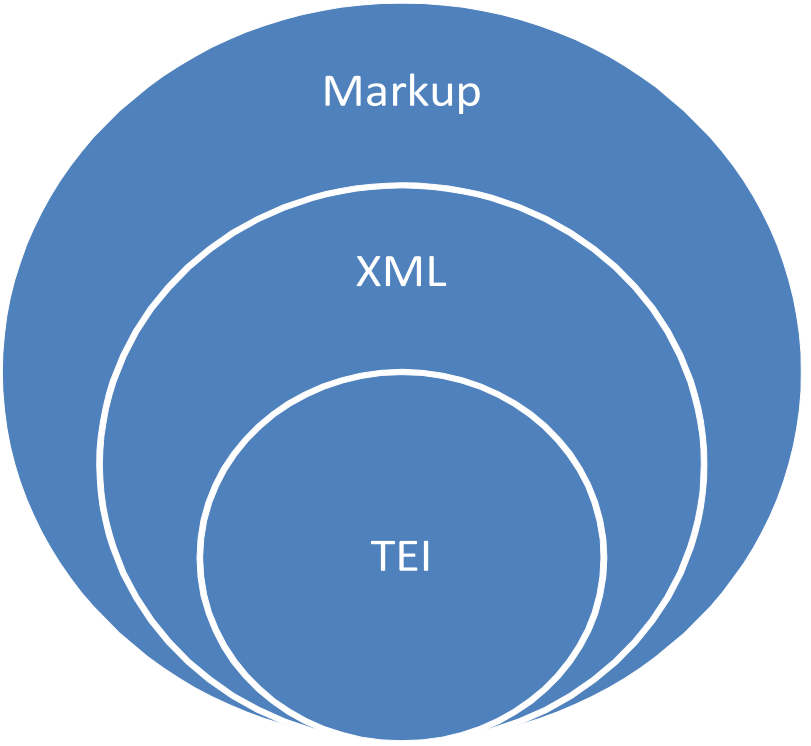
DARIAH-DE Tutorial *Digitale Textkodierung mit TEI*

Redaktion: Christof Schöch (Univ. Würzburg)

Version 1.0, 02/2014

Hinweis: Die Folien wurden unter Anpassung
eines Foliensatzes von Malte Rehbein erstellt.

Markup > XML > TEI



Was ist Markup?

Ursprung im Druckwesen:

- Anweisungen für den Setzer

“markup”

- (Subst.) = Auszeichnung, Markierung
- (Verb) = auszeichnen, markieren, kodieren

Macht Eigenschaften explizit

- benennt und/oder charakterisiert Teile der Zeichenkette
- auf formalisierte und kohärente Weise

Zwei Typen von Markup

"procedural": visuell, typographisch:

- Anweisung, wie ein Stück Text dargestellt werden soll
- Schwerpunkt auf Darstellung, Aussehen
- häufig mehrdeutig, unflexibel

"descriptive": semantisch, funktional, strukturell

- Explizieren, welche Funktion ein Stück Text hat
- Schwerpunkt auf Struktur und Bedeutung
- separat davon die Darstellung definieren
- mehr, eindeutigere Information
- Darstellung leichter anpassbar

Prozeduraler Markup

Woody Allens Urteil

Die *Recherche* ist ein tolles, aber *viel* zu langes Buch.

prozedural

```
<markup>
  <fett>Woody Allens Urteil</fett>
  <umbruch/>Die <kursiv>Recherche</kursiv> ist ein
  tolles, aber <kursiv>viel</kursiv> zu langes Buch.
</markup>
```

Deskriptiver Markup

Markup

```
<markup>
  <überschrift>Woody Allens Urteil</überschrift>
  <absatz>Die <titel>Recherche</titel> ist ein
  tolles, aber <emphase>viel</emphase> zu langes
Buch.</absatz>
</markup>
```

+

Stylesheet (Pseudo-Code)

```
überschrift =  14pt, fett, eigener Absatz
absatz      =  12pt, Blocksatz, eigener Absatz
emphase     =  12pt, kursiv, inline
kursiv      =  12pt, kursiv, inline
```

XML im Überblick

Was ist XML?

Elemente, Attribute, Werte, Strings

Prinzipien: „well-formed“

Syntax/Lexikon: „valid“

Was ist XML?

- XML = eXtensible Markup Language
- XML = Metasprache zur Definition von XML-Formaten

- XML = Standard für digitale Repräsentation von Daten
- XML = Prinzipien + Syntax

- XML = einfach (wenige, mächtige Mechanismen)
- XML = anwendungs- und plattformunabhängig

Elemente, Attribute, Werte, Strings

XML-Elemente-Attribute-Werte.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Krimi-Sammlung>
    <Krimi n="1">
        <titel>Piège pour Cendrillon</titel>
        <autor status="bekannt">
            <name>Japrisot</name>
            <vorname>Sébastien</vorname>
        </autor>
    </Krimi>
    <Krimi n="2">
        <titel>Meurtres pour mémoire</titel>
        <autor status="berühmt">
            <name>Daeninckxs</name>
            <vorname>Didier</vorname>
        </autor>
    </Krimi>
</Krimi-Sammlung>
```

well-formed vs. valid

„well-formed“ (wohlgeformt)

- Dokument entspricht den Prinzipien von XML
- die Kriterien sind immer gleich
- die Kriterien sind allgemein

„valid“ (valide)

- Dokument entspricht der Syntax und dem Lexikon eines definierten XML-Formats
- die Kriterien hängen von der jeweiligen Definition ab
- die Kriterien sind meist sehr detailliert

Kriterien für „Wohlgeformtheit“

- Prolog: XML-Version, Zeichensatz
- Nur ein Element auf oberster Ebene
- Jedes Element hat Anfangs- und Endtag
- Hierarchische Struktur: keine überlappenden Elemente
- Elemente können Unterelemente haben
- Elemente können Attribute haben
- Attribute können Werte haben
- Die Werte sind in Anführungszeichen gesetzt
- Alle Zeichen entsprechen dem ang. Zeichensatz

Wohlgeformtheit

XML-Elemente-Attribute-Werte.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Krimi-Sammlung>
    <Krimi n="1">
        <titel>Piège pour Cendrillon</titel>
        <autor status="bekannt">
            <name>Japrisot</name>
            <vorname>Sébastien</vorname>
        </autor>
    </Krimi>
    <Krimi n="2">
        <titel>Meurtres pour mémoire</titel>
        <autor status="berühmt">
            <name>Daeninckxs</name>
            <vorname>Didier</vorname>
        </autor>
    </Krimi>
</Krimi-Sammlung>
```

Noch ein paar Regeln

kann auch Kommentare enthalten

```
<!-- Kommentar →
```

kann auch „processing instructions“ enthalten

```
<?xml-stylesheet href="style.xsl"
  type="text/xsl"?>
```

Element-Namen sind „case sensitive“:

```
<name> ≠ <Name>
```

Leere Elemente können abgekürzt werden:

```
<pb></pb> = <pb/>
```

Kriterien für „Validität“

Erstens

- wohlgeformt

Zusätzlich

- Definition (Schema) vorhanden
- Dokument entspricht der Definition
- Alle notwendigen, nur erlaubte Elemente / Attribute
- Alle Werte haben eine gültige Form
- Elemente und Attribute kommen nur dort vor, wo sie auch erlaubt sind

Ein paar weitere Themen

Technologien zur Definition von XML-Formaten

- DTD (Document Type Definition)
- Schema (z.B. Relax NG)

Verwandte Technologien

- XSL / XSLT
- XPath

Mit XML definierte spezielle Markup Languages

- xHTML
- MathML, MusicXML
- TEI, MEI , CEI, ...