Abschnitt 5

Meta-Daten und Auszeichnungssprachen: XML

Wichtige Konzepte

- Daten und Meta-Daten
- Annotation
- Auszeichnungsprachen, Beispiele?
- Strukturelle vs. inhaltliche Auszeichnung

XML: Motivation und Grundlagen

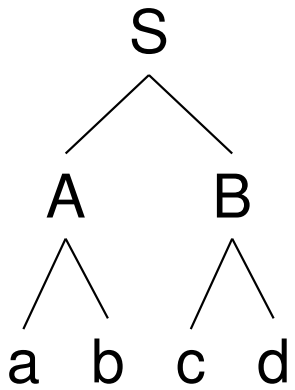
- eXtensible Mark-up Language
- **Auszeichnungssprache** zum Speichern und Austausch von Daten
- Lesbar für Menschen (möglichst **selbstbeschreibend**) *und* Maschinen (**Maschinenlesbarkeit**)
- Kein festgeschriebenes Tagset (**Erweiterbarkeit**) (Gegensatz etwa zu HTML)
- Weite Verbreitung, viele unterschiedliche Standards in unterschiedlichen Bereichen (Geisteswissenschaften: **TEI**)
- **Baumstruktur**

Beispiel zum Einstieg

```
<note>
   <from>Anna</from>
   <to>Bruno</to>
   <heading>Erinnerung</heading>
   <body>Bitte Schokolade mitbringen.</body>
</note>
```

- Beschreibt eine Notiz
- Keine Angaben zur Darstellung
- Keine Anweisungen (Unteschied zu Programmiersprachen).

Grundlegendes zu Bäumen



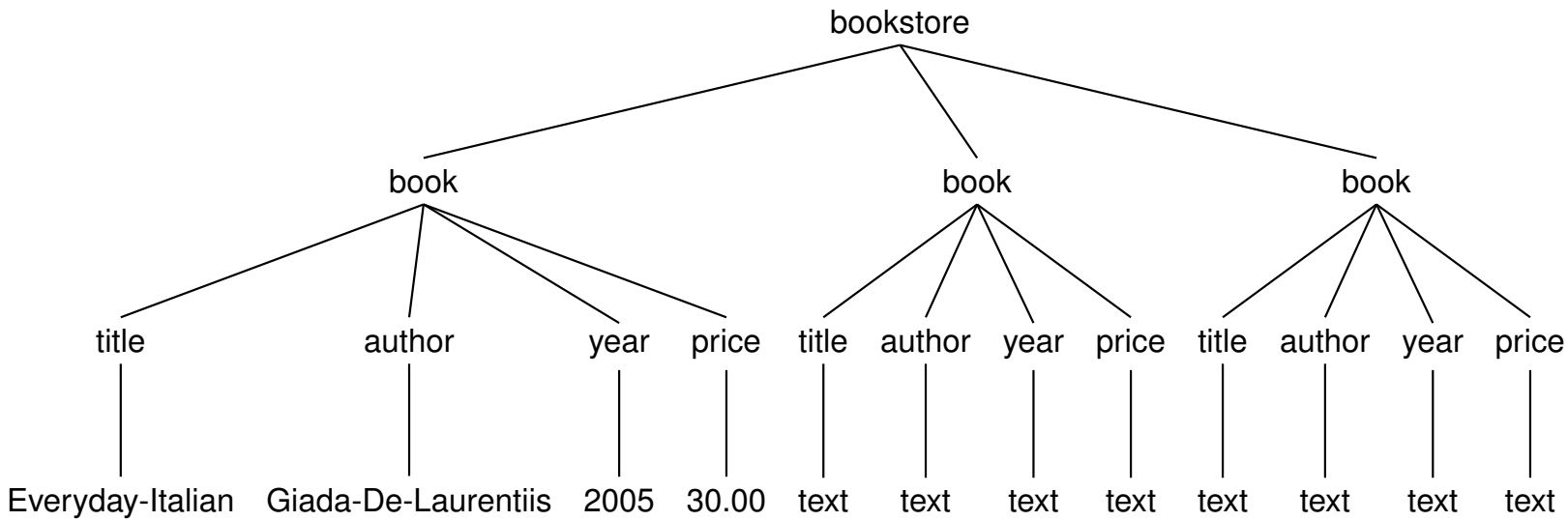
- Baum als mathematische Struktur (Graphentheorie)
- Besteht aus Knoten und Kanten
- Wichtige Beziehungen zwischen Knoten:
  - Wurzel (root): S
  - Mutterknoten (parent): S ist Mutter von A und B
  - Tochterknoten (child): A und B sind Töchter von S
  - Schwesterknoten (sibling): A und B sind Schwesterknoten
  - Blätter (leaves): a,b,c,d

XML-Dokumente als Bäume

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="web">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

http://www.w3schools.com/xml/xml_tree.asp

XML-Dokumente als Bäume



- XML-Dokumente bilden **Element-Bäume**
- Ein **XML-Element** ist ein Knoten des Element-Baums zusammen mit allen Töchtern und deren Töchtern, usw.
- Die Blätter des Element-Baums heißen **Textknoten**

Syntax

- Es muss genau ein Root-Element geben (`<bookstore>`)
- Eine optionale **XML-Deklaration** (Prologue) gibt XML-Version und Kodierung an (Standard: UTF-8)
  `<?xml version=”1.0” encoding=”UTF-8”?>`
- Alle Elemente brauchen einen **closing tag**
  `<bookstore>...<\bookstore>`
- Tags müssen richtig geschachtelt werden, also in umgekehrter Reihenfolge geschlossen werden, wie sie geöffnet worden sind
- **Attribute** werden mit Anführungszeichen umschlossen
- Keine “<” Zeichen in Textknoten
- XML-Dokumente, die diese (und einige weitere) Syntax-Regeln erfüllen, heißen **wohlgeformt**
- Wohlgeformtheit überpüfen: Browser oder externe Tools (Unix: `xmllint`)

Elemente

- Elemente bestehen aus dem Start- und dazugehörigem End-Tag und allem dazwischen
- Können Text, Attribute, oder andere Elemente beinhalten
- Dürfen **leer** sein (`<someTag></someTag>`)
- Können weitgehend frei benannt werden (keine Leerzeichen, keine Doppelpunkte,...)
- Unterscheiden Groß-/Kleinschreibung
- Können unkompliziert erweitert werden (zusätzliche Angaben stören Anwendungen zunächst nicht)

Attribute

- Beschreiben Knoten/Elemente näher (`<book category="cooking">`)
- Wert muss in (einfache oder doppelte) Anführungszeichen gesetzt werden
- Elemente können mehrere Attribute haben (`<book category="cooking" type="hardcover">`)
- **ABER**: Dieselbe Information kann auch in Töchter-Elementen gespeichert werden, was meist zu bevorzugen ist
- Allgemein gilt, dass die Information **so kleinteilig wie möglich** abgelegt werden sollte:

Schlecht:

```
<date>10-01-12</date
```

Gut:

```
<date>
    <day>12</day>
    <month>01</month>
    <year>10>year>
</date
```

Übungen zu XML (Einzelarbeit)

Stellen Sie folgenden Angaben jeweils sinnvoll als XML dar.

1. Ihr Terminkalender hat zwei Einträge: Am Dienstag, den 30. Mai, 2017 haben Sie die Übung in Digital Humanities von 17:15 bis 18:45. Danach, von 19:00 bis 20:00, essen Sie in der Abbe-Mensa zu Abend.
2. Die Buch-Triologie “Der Herr der Ringe” von J.R.R. Tolkien gliedert sich in die drei Bücher “Die Gefährten”, “Die Zwei Türme” und “Die Rückkehr des Königs”.
3. Das Modul B-GSW-12 (“Einführung in die Computerlinguistik”) wird für Studierende der Germanistik und Germanistischen Sprachwissenschaft angeboten. Das Modul dauert 2 Semester. Es wird vom Lehrstuhl für Computerlinguistik (Inhaber: Prof. Dr. Udo Hahn) angeboten und ist eine Wahlpflichtmodul.

Lösungsvorschlag (1)

```
<kalender>
  <termin>
    <was>DH Uebung</was>
    <start>
      <jahr>2017</jahr>
      <monat>Mai</monat>
      <tag>..</tag>
      <wochentag>Dienstag</wochentag>
      <zeit>
        <stunde>17</stunde>
        <minute>0</minute>
      </zeit>
    </start>
    <end>...</end>
  </termin>
  <termin>...</termin>
</kalender>
```

Lösungsvorschlag (2)

```
<buchreihe>
  <titel>Der Herr der Ringe</titel>
  <autor>Tolkien</autor>
  <buch>
    <btitel>Die Gefaehrten</btitel>
  </buch>
  <buch>...</buch>
  <buch>...</buch>
</buchreihe>
```

Unterabschnitt 1

Document Type Definition (DTD)

Grundlegendes zur DTD

- Eine DTD ist eine **optionale** Spezifikation der erlaubten XML-Elemente und deren Schachtelung
- Ein XML-Dokument, das nicht nur den allgemeinen Syntax-Regeln genügt (Wohlgeformtheit), sondern die Spezifikationen einer DTD erfüllt heißt **valide**
- Zwei technischen Umsetzungen für DTDs:
  - “Eigentliche” DTD
  - XML-Schema: selbst in XML verfasst (modernere) Alternative zur DTD

Aufbau

```
<!ELEMENT note (to,from,heading,body)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT body (#PCDATA)>
```

- Beschreibt die erlaubten Elemente und deren Inhalt
- Unterscheidet folgende Bausteine eines XML-Dokments
  - Elemente
  - Text (`#PCDATA`)
  - Attribute
  - Entities
  - #CDATA

Einbinden einer DTD in ein XML-Dokument

Intern

```
<?xml version="1.0"?>
<!DOCTYPE note [
<!ELEMENT note (to,from,heading,body)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT body (#PCDATA)>
]>
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend</body>
</note>
```

Extern

```
<?xml version="1.0"?>
<!DOCTYPE note SYSTEM "note.dtd">
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Inhalt von `note.dtd`

```
<!ELEMENT note (to,from,heading,body)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT body (#PCDATA)>
```

https://www.w3schools.com/xml/xml_dtd_intro.asp

Allgemeines zur Syntax

- Deklarationen sind jeweils in `<!...>` eingeschlossen
- Elemente werden unter der Angabe ihres Namens und des erlaubten Inhalts deklariert, z.B.:

```
<!ELEMENT note (to,from,heading,body)>
<!ELEMENT to (#PCDATA)>
```

- Bei externen DTDs wird das Root-Element zusätzlich mit `!DOCTYPE` deklariert

```
<!DOCTYPE note SYSTEM "note.dtd">
```

Element-Deklaration

- Textknoten `<!ELEMENT elementName (#PCDATA)>`
- Elemente mit beliebigen Inhalt
  `<!ELEMENT elementName ANY>`
- Elemente mit Töchterelementen:
  - Genau ein Töchterelement
    `<!ELEMENT elementName (child1)>`
  - Feste Reihenfolge an Töchterelementen
    `<!ELEMENT elementName (child1,child2,child3)>`
  - 0 oder 1 Vorkommen `<!ELEMENT elementName (child1?)>`
  - 1 bis *n* Vorkommen `<!ELEMENT elementName (child1+)>`
  - 0 bis *n* Vorkommen `<!ELEMENT elementName (child1*)>`
  - Entweder...oder
    `<!ELEMENT elementName (child1|child2)>`
  - Komplexere Kombinationen
    `<!ELEMENT elementName ((child1|child2)+,child3?)>`

Übungen zur DTD (Gruppenarbeit)

Im folgenden werden verschiedene Strukturen beschrieben, die sich im konkreten Fall durch ein XML-Dokument darstellen ließen. Stellen Sie jeweils eine DTD auf, die diese allgemeine Struktur beschreibt.

1. Der Tageswetterbericht besteht aus den Vorhersagen für den Vormittag, den Nachmittag und die Nacht. Für jede dieser Tageszeiten, wird die Temperatur, die Niederschlagswahrscheinlichkeit und die Windrichtung angegeben.
2. Ein Fachschaftsrat setzt sich aus mehreren Mitgliedern zusammen. Jedes dieser Mitglieder ist zu einem bestimmten Zeitpunkt in den FSR eingetreten. Es gibt gewählte und nicht-gewählte Mitglieder. Darüber hinaus nehmen manche Mitglieder bestimmte Funktionen innerhalb des FSRs ein. Solche Funktionen sind etwa die des Vorsitzendenden oder des Schatzmeisters. Ein Mitglied kann mehrere Funktionen haben.
3. Eine Bibliothek setzt sich aus vielen verschiedenen Abteilungen zusammen. Jeder Abteilung ist ein Abteilungsleiter zugeordnet, der über die unterschiedlichen Bücher, die wiederum nach Fachgebiet unterschieden werden, bestimmt. Jedes Buch hat eine Signatur.