Digital Humanities: Übung 3

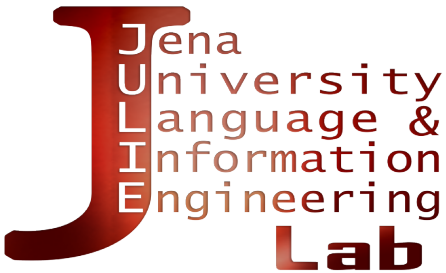
Meta-Daten und Auszeichnungssprachen: XML

Sven Büchel

Friedrich-Schiller-Universität Jena
Philosophische Fakultät
Institut für Germanistische Sprachwissenschaft
Lehrstuhl für Computerlinguistik
(JULIE Lab)
Sommersemester 2017

http://www.julielab.de

06. Juni 2017



Agenda

Aufgaben von letzter Woche

Meta-Daten und Auszeichnungssprachen: XML

Aufgabe zu nächster Woche

Aufgaben von letzter Woche

Aufgabe 2.1: Reguläre Ausdrücke

Geben Sie einen Regulären Ausdruck zum Auffinden von akademischen Titeln an. Der Ausdruck muss mindestens die folgenden Beispiele erfassen.

- Dr.
- Prof. Dr.
- Dr. h.c.
- Dr. phil.
- Prof. Dr. med.
- Prof. emer.
- PD Dr.
- PD Dr. rer. nat.

Der Ausdruck sollte keine sich wiederholenden Teile enthalten (also nicht einfach alles durch Disjunktion lösen)!

Aufgabe 2.1: Reguläre Ausdrücke (Lösung)

(((Prof\.|PD)␣)?Dr\.(␣(h\.c\.|med\.|phil\.|rer\. nat\.))?|Prof\. emer\.)

Aufgabe 2.2: Suche im DTA

1. Geben sie die Adjektive an, die in Werken des Zeitraums 1700-1750 vor dem exakten Wort *„Gevatter“* stehen.
2. Welches ist das älteste Werk im DTA, das eine Wortform des Lemmas *Eisenbahn* am Anfang eines Satzes enthält?

Geben Sie jeweils auch die dafür nötige Suchanfrage an!

Aufgabe 2.2: Suche im DTA (Lösung)

- "$p=ADJA @Gevatter" #less_by_date[1700,1750]
- Eisenbahn with $.=0 #less_by_date

Meta-Daten und Auszeichnungssprachen: XML

Wichtige Konzepte

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

XML: Motivation und Grundlagen

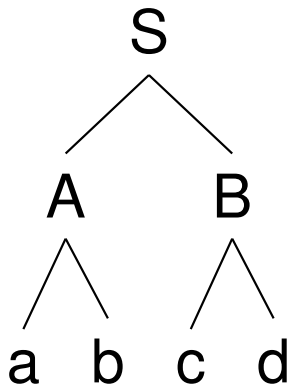
- eXtensible Mark-up Language
- **Auszeichnungssprache** zum Speichern und Austausch von Daten
- Lesbar für Menschen (möglichst **selbstbeschreibend**) *und* Maschinen (**Maschinenlesbarkeit**)
- Kein festgeschriebenes Tagset (**Erweiterbarkeit**) (Gegensatz etwa zu HTML)
- Weite Verbreitung, viele unterschiedliche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



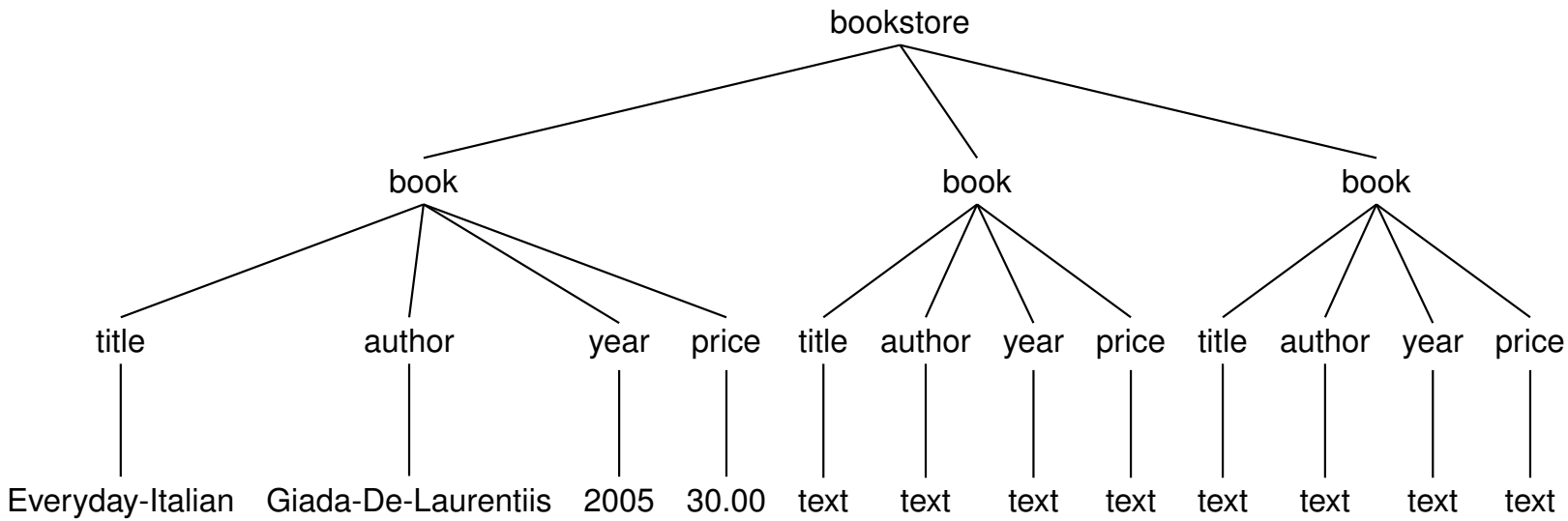
- Baum als mathematische Struktur (Graphentheorie)
- Besteht aus Knoten und Kanten
- Wichtige Beziehungen zwischen Knoten:
  - Wurzel (root): S
  - Mutterknoten (parent): S ist Mutter von A und B
  - Tochterknoten (child): A und B sind Töchter von S
  - Schwesterknoten (sibling): A und B sind Schwesterknoten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

Aufgabe zu nächster Woche

Formalia

- Abgabe bis Montag, den 12.6., 23:59 Uhr
- Per Email an `sven.buechel@uni-jena.de`
- Im PDF-Format
- Vorname, Name, Matrikelnummer und Veranstaltungsname auf dem Blatt
- Gruppenarbeit ausdrücklich erlaubt! Trotzdem separate Abgabe.

Aufgabe 3.1

Laden Sie sich folgenden Forschungsartikel herunter:

> *Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, T., . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182.*

Erstellen Sie aus den ersten drei Einträgen des Literaturverzeichnisses dieses Papiers eine verkleinerte Bibliographie, wobei Sie die Daten mit XML auszeichnen. Verwenden Sie dabei z.B. folgende Element-Namen: bibliography, publication (mit Attribut book, article,...), author, firstname, lastname. Verwenden Sie weitere selbstgewählte Element-Namen, sodass Sie die Daten möglichst kleinteilig auszeichnen können. Dabei sollen die untersten Knoten im Element-Baum möglichst *nur eine* Art von Informationen beinhalten (also etwa *nicht* Vor- und Nachname des Autoren zusammenfassen). Achten Sie bitte auf sinnvolle Formatierung bzw. Einrückung.