tutego™
YOUR TRAINING COMPANY

Regeln für XML-Dokumente

Übersicht: Regeln für XML-Dateien

- XML-Deklaration und Verarbeitungsanweisungen
- Die Dokumenttyp-Deklaration (DTD)
- Regeln für Tags, Attribute, Werte und Kommentare
- Wohlgeformt und gültig
- Baumstruktur und Knoten einer XML-Datei
- Zeichen, Zeichensätze und nicht interpretierte Abschnitte

Markup-Typen

- XML-Dokumente bestehen aus Markups mit folgenden Typen
  - Element: Speichereinheit, z. B. <artikel>Hose</artikel>
  - Kommentar: <!-- Kommentar -->
  - CDATA schützt Textblöcke
  - Entity: Abkkürzung für Textbausteine
  - ProcessInstruction/Notation: Zusätzliche Anweisungen und Definition von Dateitypen

Prolog

- Durch die Auszeichnung am Anfang wird der Bezug zu XML hergestellt:

```
<?xml version="1.0"?>
```

- Mit dem Attribut encoding wird der verwendete Zeichensatz angegeben:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

- Das Attribut standalone sagt dem Parser, ob eine externe DTD verwendet wird. Mögliche Werte sind yes und no. Der Standard ist no.

```
<?xml version="1.0" standalone="yes"?>
```

Kodierungen

- Als Belegungen des Encoding-Namens sind beliebige Zeichensätze zugelassen.
- Weit verbreitet sind folgende Deklarationen
  - ISO-8859 (extended ASCII)-Familie
  - UTF-8
  - UTF-16

```
<?xml version="1.0" encoding="Shift_JIS" standalone="yes"?>
<kougi>
  <title begin="二千一年三月二十一日三時三十分＋九時">選挙義務講義ＸＭＬ</title>
  <organization>アウグスブルグ大学コンピュータ・サイエンス過分、工学と家政と形成の大学</organization>
</kougi>
```

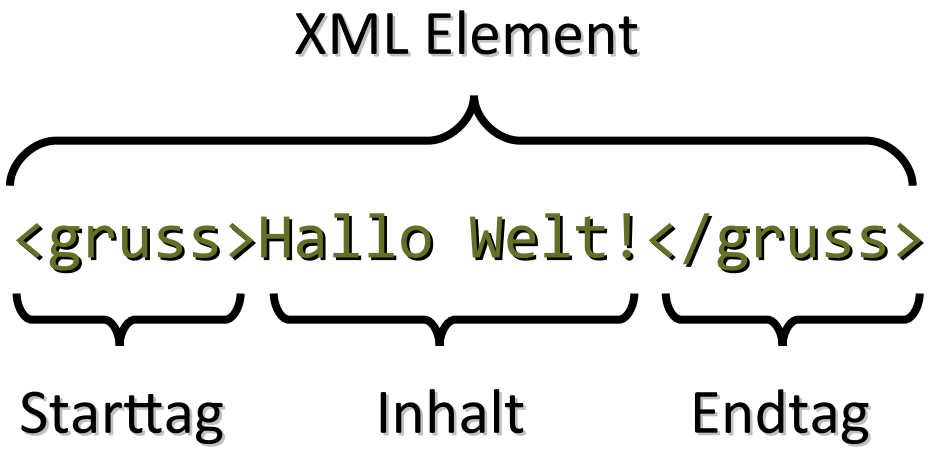
Kommentare

- Kommentare werden durch die Zeichen <!-- und --> markiert.
  - Ein Kommentar darf nicht mit ---> abgeschlossen werden und -- darf nicht vorkommen.
- Sie dürfen an jeder beliebigen Stelle im XML-Dokument stehen, nur nicht innerhalb der Tag-Namen.

```
<!-- Hier liest kein Parser -->                    ✔

<!-- - - - Hier liest kein Parser - - - -->        ✔

<!-- **** Hier liest kein Parser **** -->          ✔

<!-------- Hier liest kein Parser --------->       ✘

<!-- -- -- Hier liest kein Parser -- -- -->        ✘
```

Elemente einer XML Datei

- Jedes Element besteht aus einem öffnenden und einem schließendem Tag, genannt Start-Tag und End-Tag.
- Dazwischen befindet sich der Elementinhalt (Content).



- Sprachlich mischt man gerne den Begriff Element und Tagname.

Leeres Element

- Falls ein Element keinen Inhalt hat, kann abgekürzt werden.
- In HTML steht z. B.
 für einen Zeilenvorschub. Dieses Element hat keinen Inhalt!
- Statt
</br> schreibt man für dieses leere Element
.

Die Dokumenttyp-Deklaration

- Mit dieser Deklaration wird eine Verbindung zwischen den XML-Daten und der DTD hergestellt.
- Die DTD kann in einer Datei stehen oder Bestandteil der XML-Datei sein.
- Öffentliche DTD-Deklarationen verweisen auf DTDs, die durch einen Standard festgelegt sind.
- Eine Kombination von interner und externer DTD ist möglich.

Besondere Zeichen

- Bestimmte Zeichen haben in XML eine festgelegte Bedeutung und müssen umkodiert werden:
  - < für <
  - & für &
- Üblicherweise werden auch folgende Zeichen umkodiert:
  - > für >
  - " für "
  - ' für '

CDATA

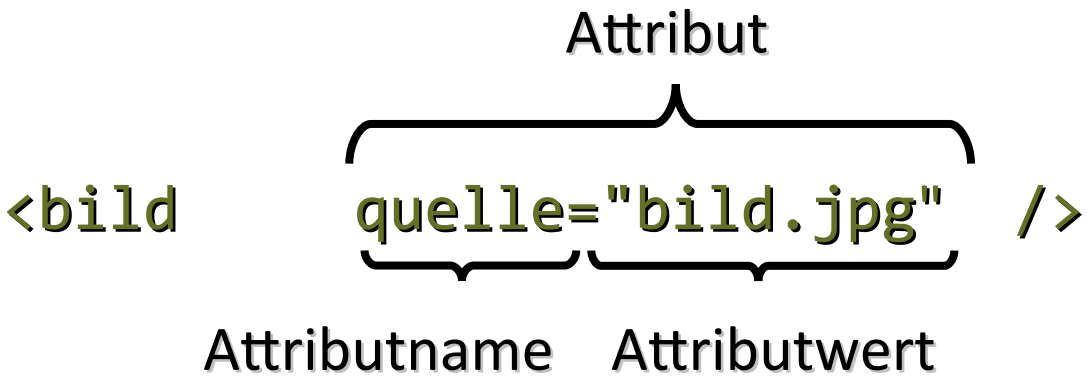
- CDATA schützt Textblöcke, die sonst als Markup interpretiert würden. Sie können nicht ineinander verschachtelt werden.

```
<![CDATA[
  <gruss> Hallo, Welt! </gruss>
]]>
```

- Die Zeichen ']]' sind in CDATA verboten.

Attribute

- Neben dem Inhalt kann jedes Element beliebig viele Attribute enthalten.
- Attribute werden in dem öffnenden Tag eingetragen.
- Jedes Attribut muß einen Wert in einfachen oder doppelten Anführungszeichen besitzen.



Elementinhalt - Attributwert

- Der Designer einer XML-Datei muss entscheiden, ob der Inhalt als Attribut oder als Elementinhalt gespeichert wird.
  - Ein Element kann nur einen Textinhalt aber viele Attribute und Unterelemente enthalten.
  - Attribute werden oft für Parameter und untergeordnete Inhalte verwendet.
- Im Zweifelsfall wird ein neues Tag untergeordnet.

Regeln für Tags, Attribute

- XML unterscheidet streng zwischen Groß- und Kleinschreibung.
- Elementnamen
  - sollten nicht mit der Zeichenkette xml beginnen
  - dürfen keine Leer- oder Gleichheitszeichen enthalten
  - beginnen mit einem Buchstaben oder _
  - dürfen keinen Doppelpunkt enthalten

Verarbeitungsanweisungen

- Anweisungen an die Software, die eine XML-Datei verarbeitet, werden mit `?` zwischen öffnendem und schließendem Tag markiert.
  - Diese Verarbeitungsanweisungen heißen Processing Instructions (PI).
  - Hinweis: Die erste Zeile sieht zwar so aus wie eine PI, ist aber keine!
- Verarbeitungsanweisungen, wie das Einbinden eines Stylesheets, sind durch Standards festgelegt und werden von vielen Anwendungen unterstützt.

```
<?apache include file="fusszeile.html" ?>
```

Wohlgeformt und Gültig

- Eine XML-Datei ist gültig, wenn sie wohlgeformt ist und der Struktur einer formalen Beschreibung folgt.
- Diese formale Beschreibung kann eine DTD sein.
- Andere Möglichkeiten zur Strukturbeschreibung:
  - Schema: Eine XML-Applikation in der komplexe Datentypen und deren Verschachtelung beschrieben werden.
  - RELAX NG: **RE**gular **LA**nguage **X**ML Strukturbeschreibung durch reguläre Ausdrücke. *http://www.relaxng.org/*

Baumstruktur, Knoten einer XML-Datei

- Jede XML-Datei entspricht einer Baumstruktur:
- Es gibt genau ein Wurzelelement.
- Die Verschachtelung der Elemente bildet die Baumstruktur auf Elementebene.
- Attribute sind den Elementen als Knoten untergeordnet.
- Elementwerte und die Werte der Attribute bilden die Blätter des Baums.

Zusammenfassung

- Durch die XML-Deklaration wird am Anfang der Datei festgelegt, dass es sich um eine XML-Datei handelt.
- Verarbeitungsanweisungen und Dokumenttyp-Deklarationen haben eine spezielle Syntax.
- Inhalte können zwischen Elementen oder als Attribute in der XML-Datei stehen.
- Kommentare helfen Lesern der XML-Datei.
- XML legt Regeln für Elemente und Attribute fest.
- Durch eine DTD wird die Struktur beschrieben.
- Jede XML-Datei besitzt eine eindeutige Baumstruktur.