Wohlgeformtheit von XML-Dokumenten

Mikrostruktur

Auch wenn XML dem Kodierenden im Vergleich zu HTML eine freiere Textauszeichnung erlaubt, muss auch die XML-Kodierung bestimmten Regeln folgen, um als wohlgeformt gelten zu können.

Grob betrachtet besteht ein XML-Dokument aus **Tags** und **Inhalt**. Der Inhalt, gewissermaßen der pure Text, wird durch Tags, erkennbar an der Klammerschreibweise (<...>), mit Informationen angereichert:

```
<nomen> Auszeichnungssprache </nomen>
```

Der Inhalt Auszeichnungssprache wird hier mithilfe der Tags als *nomen* ausgezeichnet. Tags müssen stets als Paar auftreten, als eine Kombination von **St art-Tag** und **End-Tag**, in unserem Beispiel *<nomen>* und *</nomen>*. Abgesehen von dem Slash, der das End-Tag markiert, müssen Start- und End-Tag identisch sein, auch im Hinblick auf die Großschreibung. Das Starttag *<nomen>* in Kombination mit dem Endtag *</Nomen>* bilden keine wohlgeformte XML-Struktur und würden beim Parsen des Dokuments eine Fehlermeldung hervorrufen. Anfangs- und End-Tag können alternativ auch in einem einzigen Klammerpaar zusammengefasst werden:

```
<pause></pause>

<pause/>

<pause />
```

Die beiden alternativen Schreibweisen eignen sich für Tags, deren Anfangs- und End-Tag direkt aufeinander folgen würden und welche somit keinen Inhalt rahmen, wie zum Beispiel in *<satz> Das ist <fehlender_artikel/> Auszeichnungssprache </satz>* oder *<satz> Das ist eine <pause/> Auszeichnungssprache </satz>*.

Start- und End-Tag bilden Anfang und Ende eines **Element**s. Elemente können andere Elemente enthalten und auf diese Weise eine hierarchische Struktur schaffen. *<wort><nomen>Auszeichnungssprache</nomen></wort>* ließe sich folgendermaßen visualisieren:

```
<wort>
        <nomen> Auszeichnungssprache </nomen>
</wort>
```

Das Element *<nomen>* befindet sich innerhalb des Elements *<wort>* und liegt folglich eine Ebene tiefer in der Elementhierarchie. Elemente können zwar geschachtelt sein, dürfen sich jedoch nicht überkreuzen. Während eine Struktur wie *<wort><nomen>Auszeichnungssprache</nomen></wort>* als wohlgeformt gilt, steht *<wort><nomen>Auszeichnungssprache</wort></nomen>* im Widerspruch zu den Regeln der XML-Struktur.

```
<wort>
        <nomen> Auszeichnungssprache </nomen>
</wort>
```

Im Vergleich zu der nicht-wohlgeformten Struktur:

```
<wort>
        <nomen> Auszeichnungssprache
</wort>
        </nomen>
```

Einem Element können zudem Attribute und Werte zugeordnet werden, um mitgetragene Informationen zu präzisieren und unnötige Hierarchietiefen zu vermeiden. *<wort><nomen>Auszeichnungssprache</nomen></wort>* ließe sich ohne Informationsverlust auch als *<wort wortart="nomen" >Auszeichnungssprache</wort>* darstellen. Dabei bildet *wortart* das Attribut, das in die Klammer durch eine Leerstelle getrennt auf den Elementnamen im Anfangs-Tag folgt. Der durch ein = mit dem Attribut verbundene Wert wird in Anführungsstriche gesetzt. In der Regel kann ein Element beliebig viele Attribute/Werte fassen.

```
<wort wortart="nomen" genus="f" wortbildung="kompositum"> Auszeichnungssprache </wort>
```

Makrostruktur

Neben der Mikrostruktur ist auch die Makrostruktur eines XML-Dokuments für dessen Wohlgeformtheit von Relevanz.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes ?>
<beispiel>
        <titel> Ein Satz </titel>
        <satz>
                <subjekt>
                        <pronomen> Ich </pronomen>
                      </subjekt>
                <prädikat>
                              <verb> lerne </verb>
                </prädikat>
                <objekt>
                        <artikel>eine</artikel>
                        <nomen>Auszeichnungssprache</nomen>
                </objekt>
        </satz>
</beispiel>
```

Jedes XML-Dokument beginnt mit einem Kopf, welcher aus einer XML-Deklaration und einer fakultativen Dokumenttypdeklaration besteht. In dem vorliegenden Beispiel enthält der Kopf lediglich eine XML-Deklaration, welche anhand der Form *<?xml ... ?>* zu erkennen ist.

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

Innerhalb der spitzen Klammern werden den drei Attributen *version*, *encoding* und *standalone* Werte zugeordnet. *Version* meint die XML-Version, mit der gearbeitet wird. Hier sollte im Normalfall 1.0 gewählt werden. *Encoding* definiert, welcher Zeichensatz in dem Dokument zugelassen ist. Die Norm ISO-8859-1 ist dabei für die meisten romanischen Sprachen ausreichend. Ausnahmen bilden das Französische und das Rumänische, die zwar größtenteils, aber nicht komplett durch die Norm gedeckt sind. So enthält ISO-8859-1 weder *œ* (z.B. *cœur*) noch das überaus seltene *ÿ* (in Eigennamen). Hier können alternativ die Normen ISO-8859-15 oder ISO-8859-16 verwendet werden. Auch fehlen in dem ISO-8859-1-Satz einige Diakritika des Rumänischen, die in ISO-8859-2 einbegriffen sind.
Das Attribut *standalone* geht mit der Dokumenttypdeklaration einher. Wird *standalone* auf *yes* gesetzt, befindet sich diese entweder im Dokument selbst oder sie wird gar nicht gegeben. *no* bedeutet hingegen, dass die Dokumenttypdeklaration extern zu suchen ist.

Die Dokumenttypdeklaration, erkennbar an *<!DOCTYPE ... >*, soll an dieser Stelle vorerst ausgeklammert werden. Sie definiert weitere Beschränkungen (bzw. gibt an, wo sie definiert werden), denen das Dokument unterliegen muss. Erfüllt ein Dokument die dort vorgegebenen Kriterien, kann es als *valide* bezeichnet werden.

Ein XML-Dokument gilt nur als Wohlgeformt, wenn diese Attribute und Werte der Norm entsprechend genannt werden. Der Rest des Dokuments folgt ebenfalls einigen Regeln der Wohlgeformtheit.

Betrachtet man die *pronomen*, *verb*, *artikel* und *nomen* übergeordneten Elemente *titel* und *satz*, stellt man fest, dass sie nicht die erste Hierarchieebene bilden, sondern dem Element *beispiel* untergeordnet sind. Dies resultiert aus einer weiteren Regel für die Wohlgeformtheit von XML-Dokumenten, nach welcher die erste Hierarchieebene (klammert man den Kopf aus) aus genau einem Element bestehen muss. Dieses Element wird als **Wurzelelement** bezeichnet und rahmt das gesamte sich unter dem Kopf befindende XML-Dokument. Während also jede weitere Hierarchieebene beliebig viele Elemente enthalten kann, besteht die erste Ebene aus genau einem Element, dem Wurzelelement. Untergeordnete Elemente werden als **Kinder** (**children**) bezeichnet, übergeordnete als **Eltern** (**parents**). Elemente auf der zweiten Hierarchieebene sind Kinder des Wurzelelements und Eltern der Elemente der dritten Hierarchieebene. Elemente der dritten Ebene sind Kinder der Elemente der zweiten Ebene und Eltern der Elemente der vierten Ebene, u.s.w. .

Die Wohlgeformtheit eines XML-Dokuments ist gegeben, wenn:

- die Regeln für die XML-Syntax befolgt werden. D.h. auch, dass:
  - der Kopf (XML-Deklaration und Dokumenttypdeklaration) dem XML-Standard entspricht,
  - der Zeichensatz des Dokuments dem im Kopf definierten entspricht,
  - das Dokument ein einziges Wurzelelement besitzt,
  - jedem Anfangstag auch ein Endtag zugeordnet ist,
  - sich die Elemente nicht überlappen,
  - die Groß- und Kleinschreibung beachtet wird,
  - Attributwerte in Anführungszeichen gesetzt sind.

Unbekanntes Makro: 'scrollbar'