Vorlesung / Übungen

Multimedia Technologie II

Prof. Dr. Michael Frank / Prof. Dr. Klaus Hering

Sommersemester 2004

HTWK Leipzig, FB IMN

Für die externe Vorhaltung der DTD werden sämtliche zwischen den eckigen Klammern der DOCTYPE-Deklaration befindliche Deklarationen in ein separates File (etwa katalog.dtd) geschrieben. Unter der Annahme, dass dieses File im gleichen Verzeichnis wie das betreffende XML-File steht, sieht die modifizierte DOCTYPE-Deklaration wie folgt aus:

```
<!DOCTYPE Katalog SYSTEM “katalog.dtd“>
```

Parser, die neben der Wohlgeformtheit auch die Gültigkeit eines Dokuments in Bezug auf eine DTD untersuchen:

XML4J (IBM) **ProjectX** (Sun) **MSXML** (Microsoft)
XML Parser for Java (Oracle)

Web-based XML validation

STG's XML 1.0 Reference Validator

Scholarly Technology Group,
Brown University, Providence, USA

http://www.stg.brown.edu/service/xmlvalid/

XML Well-Formedness Checker and Validator

R. Tobin, Language Technology Group (LTG),
University of Edinburgh, UK

http://www.cogsci.ed.ac.uk/%7Erichard/xml-check.html

Die Elementstruktur eines XML-Dokuments kann durch *Elementtyp-Deklarationen* und durch *Attributlisten-Deklarationen* eingeschränkt werden.

Elementtyp-Deklarationen

⇒ Einschränkung des Elementinhaltes (4 Fälle)

<!ELEMENT *name* (EMPTY | ANY | *mixed_content* | *element_content*)>

EMPTY	Elementinhalt leer
ANY	Elementinhalt beliebig (aber wohlgeformt!)
MIXED CONTENT	Elementinhalt besteht aus reinem Text oder letzterem versetzt mit Elementen aus einer vorgegebenen Menge
ELEMENT CONTENT	Elementinhalt besteht lediglich aus Elementen, wobei die Elementanordnung durch eine Grammatik beschrieben ist

Spezialfälle

```
<?xml version="1.0"?>
<!DOCTYPE Dok [
<!ELEMENT Dok ANY>
<!ELEMENT A  EMPTY>
]>
<Dok>
a<A/>b
</Dok>
```

gültig

```
<?xml version="1.0"?>
<!DOCTYPE Dok [
<!ELEMENT Dok ANY>
<!ELEMENT A EMPTY>
]>
<Dok>
<A>ab</A>
</Dok>
```

ungültig

Ausdrucksmittel zur Syntaxbeschreibung

,	strikte Reihenfolge	(Infix)
?	Optionales Vorkommen	(Postfix)
+	ein- oder mehrmaliges Vorkommen	(Postfix)
*	kein- oder mehrmaliges Vorkommen	(Postfix)
|	Auswahl genau eines Vorkommens	(Infix)
()	Gruppierung (zu komplexem Vorkommen)	(Prä-/ Postfix)

Wir benutzen diese Ausdrucksmittel im folgenden auf zwei Ebenen:

- Beschreibung der Syntax von Elementtyp-Deklarationen
- innerhalb von Elementtyp-Deklarationen zur Beschreibung der Elementstruktur

Grammatik

Wir verstehen unter einer *Grammatik G* ein Mittel zur Beschreibung einer *formalen Sprache L(G)*. Eine formale Sprache verkörpert eine Menge von Wörtern über einem Alphabet *T* von *Terminalsymbolen*.

G = (W, N, T, R)

N ... Menge von Nichtterminalsymbolen (Hilfssymbolen)

W ... Sprachwurzel (ausgezeichnetes Nichtterminalsymbol)

T ... Menge von Terminalsymbolen

R ... Menge von Regeln der Form

Wort ::= Beschreibung einer Wortmenge

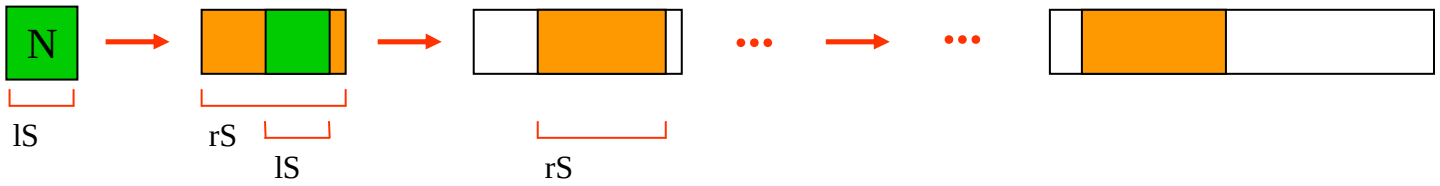
linke Seite rechte Seite

Treten als linke Regelseiten lediglich Nichtterminalsymbole auf, sprechen wir von einer *kontextfreien Grammatik*.

Zur Beschreibung der Wortmenge auf der rechten Seite einer Regel lassen wir die Benutzung der oben beschriebenen technischen Zeichen „,“ , „?“, „+“ , „*“ , „|“ , „(„ und „)“ zu.

Interpretation einer Regel:
Ein Wort aus der zur rechten Regelseite gehörenden Wortmenge kann das die linke Regelseite verkörpernde Wort ersetzen.

Ableitung aus der Sprachwurzel N:
Folge von Ersetzungen entsprechend der Regelmenge R ausgehend von N



Die durch G beschriebene Sprache L(G) ist die Menge der aus N ableitbaren Wörter, welche nur aus Terminalsymbolen bestehen.

Grammatik für die Kategorie *mixed_content*

mixed_content ::= (#PCDATA) |

(#PCDATA (| *name*)*)*

Die erste Auswahlmöglichkeit beschreibt den Elementinhalt als reinen Text. Die zweite Möglichkeit läßt im Text Vorkommen von (beliebig vielen) Elementen entsprechend der auftretenden Elementnamen zu (ohne Einschränkungsmöglichkeit der Anordnung).
Zusatzbedingung: Ein Elementname darf nicht mehrfach vorkommen.

Beispieldeklaration: *mixed_content*

```
<!ELEMENT Dok ( #PCDATA | A | B )* >
```

Gültiges Dokument

```
<?xml version="1.0"?>
<!DOCTYPE Dok [
<!ELEMENT Dok (#PCDATA | A | B)*>
<!ELEMENT A (#PCDATA)>
<!ELEMENT B EMPTY>
<!ELEMENT C (#PCDATA)>
]>
<Dok>
abc<A>de</A><A/>f<B/>gh
</Dok>
```

Einfügen von <C>h</C> führt zur Ungültigkeit

Grammatik für die Kategorie *element_content*

element_content ::= (*choice* | *sequence*) (? | * | +) ?

choice ::= (*object* (| *object*)*)

sequence ::= (*object* (, *object*)*)

object ::= (*name* | *choice* | *sequence*) (? | * | +) ?

Im Unterschied zur Kategorie *mixed_content* sind hier Anzahl- und Anordnungsbeschränkungen für im Inhalt vorkommende Elemente formulierbar.

Beispieldeklaration:

element_content

<!ELEMENT Dok ((A , B*) | (B , A)+)>