XML – Merkblatt 1 – Grundlagen

Einführung Die Hyper Text Markup Language (HTML) ist die Sprache des World Wide Webs (WWW). Es ist eine Seitenbeschreibungssprache, die Text visuell repräsentiert und von einem Web-Browser angezeigt werden kann. Texte werden mit *Auszeichnungen* (markups) formatiert. Die Auszeichnungen können nicht frei gewählt werden. Sie sind im HTML-Standard definiert. Die Auszeichnungen stellen Klammern bestehend aus einem Start-Tag und einem End-Tag dar. `<p>...</p>` (paragraph) weist zum Beispiel den Browser an, den Textinhalt (`...`) in einem Absatz darzustellen.

```
<html>
  <head></head>
  <body>
    <h1>Bibliography</h1>
    <p><i>Foundations of Databases</i>
      Abiteboul, Hull, Vianu<br/>
      Addison Wesley, 1995
    </p>
    <p><i>Data on the Web</i>
      Abiteboul, Buneman, Suciu<br/>
      Morgan Kaufmann, 1999
    </p>
  </body>
</html>
```

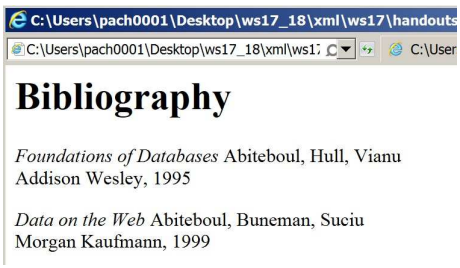

Abbildung 1: Links ein HTML-Beispiel mit einer Überschrift (h1), zwei Absätzen (p) und Text, der teilweise kursiv (i) gesetzt ist. Rechts davon dessen Darstellung in einem Web-Browser.

HTML eignet sich nicht gut für den automatischen Austausch von Daten zwischen Rechnern. XML (eXtensible Markup Language) wurde als semi-strukturiertes Datenmodell entwickelt, das sowohl für den Menschen lesbar als auch für den Rechner leichter zu verarbeiten ist. XML 1.0 ist derzeit die am weitesten verbreitetste und genutzte Version. Struktur und Auszeichnungen von XML sind erweiterbar. XHTML ist zum Beispiel eine XML-Variante von HTML mit etwas strikteren Syntaxregeln. Auf Basis des XML-Standards gibt es eine Fülle weitere XML-Sprachen für spezifische Anwendungszwecke: XML Schema, XSL, XPath und mehr. Sowohl XML als auch HTML sind vereinfachte Versionen des SGML-Standards (Standarized General Markup Language).

Semi-strukturierte Daten bestehen aus Folgen von Objekten, die wieder in Objekte oder Attribute unterteilt werden können. Die Attribute enthalten Werte eines bekannten Datentyps. In XML sind die Daten hierarchisch unterteilt, sie lassen sich deswegen auch als Baum darstellen. Diese XML-Daten werden als XML-Dokument bezeichnet. XML-Dokumente wer-

```
<entry>
  <name>
    <fn>Jean</fn>
    <ln>Doe</ln>
  </name>
  <work>
    INRIA
    <adress>
      <city>Cachan</city>
      <zip>94235</zip>
    </adress>
    <email>j@inria.fr</email>
  </work>
  <purpose>like to teach</purpose>
</entry>
```

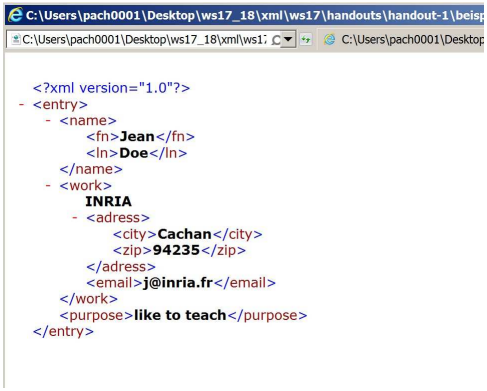

Abbildung 2: Links ein XML-Dokument mit frei gewählten Auszeichnungen. Rechts davon dessen Darstellung in einem Web-Browser.

den üblicherweise in serialisierter Form in Textdateien gespeichert. Die de-serialisierte Repräsentation als Baum wird normalerweise intern im Hauptspeicher eines Rechners verwendet.

Syntax von XML Dokumenten XML unterstützt Unicode für die Repräsentation von Zeichen. XML-Dokumente bestehen aus Elementen und Text. Ein Element besteht aus Start- und End-Tag gleichen Namens und dem Inhalt dazwischen. Start-Tags können Attribute bestehend aus Namen und Wert enthalten. Der Name muss eindeutig pro Element sein. Der Wert eines Attributes wird entweder in einfachen oder in doppelten Hochkommas gesetzt. Attribute haben keine definierte Reihenfolge.

```
<name attribut='wert' ...> Inhalt </name>
```

Ein leeres Element `<a></a>` kann als `<a/>` abgekürzt werden. Die Tag-Namen sind wie alle XML-Bezeichner frei wählbare Folgen von Unicode Zeichen, Ziffern, Doppelpunkt (`:`), Bindestrich (`-`), Unterstrich (`_`), Punkt (`.`) und Mittelpunkt (·). Ein Tag-Name darf nicht mit einer Ziffer beginnen. Tag-Namen sind case-sensitive. Der Doppelpunkt sollte nicht verwendet werden, er ist für Namensräume reserviert Ebenso sollte ein Tag-Name nicht mit der Zeichenfolge `xml` beginnen, egal in welcher Mischung aus Groß- oder Kleinbuchstaben. Ein Element kann als Inhalt selbst eine Folge von Elementen und Text enthalten.

In XML können für den menschlichen Leser Kommentare (`<!-- wird ignoriert -->`) überall außer in den Tags angegeben werden. Der Kommentar kann sich über mehrere Zeile erstrecken, darf aber nicht geschachtelt sein.

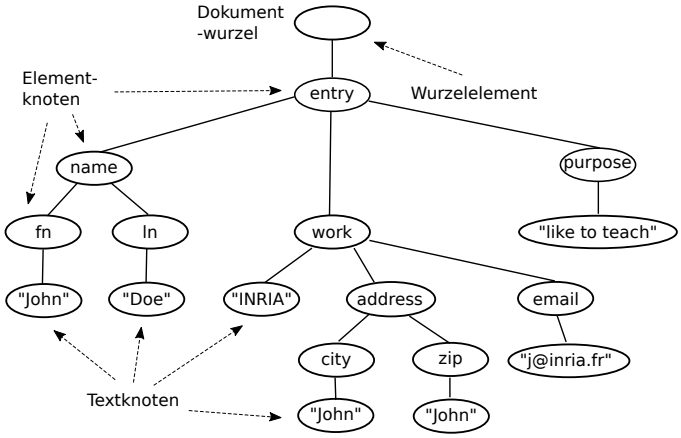

Abbildung 3: Baumdarstellung des XML-Dokuments aus Abbildung 2.

```
<document/>
```

```
<document>
  Hello World!
</document>
```

```
<document>
  <salutation>
    Hello World!
  </salutation>
</document>
```

```
<?xml version="1.0" encoding="utf-8"?>
<document>
  <!-- a comment -->
  <salutation color="blue">
    Hello World!
  </salutation>
  <salutation color="red">
    <![CDATA[<Hello Moon>!]]>
  </salutation>
</document>
```

Abbildung 4: Vier verschiedene wohlgeformte XML-Dokumente.

`&` und `<` dürfen nicht als Textinhalt in Auszeichnungen verwendet werden, stattdessen können die aus HTML bekannten Fluchtsymbole `&` und `<` benutzt werden. Dies gilt auch für Attributwerte. Attributwerte, Kommentare und CDATA-Abschnitte zählen nicht als Auszeichnungen.

Mit einer Zeichenreferenz `&#DDDD;` oder `&#xFFFF;` lassen sich beliebig Unicode-Zeichen im Text angeben, wobei `DDDD` dessen Codierung als Dezimalzahl und `FFFF` als Hexadezimalzahl. Die Anzahl Ziffern der Zahl ist nicht beschränkt. `
` und `
` stellen beide den Zeilenumbruch (line feed) dar.

Ein CDATA-Abschnitt darf überall in einem Element verwendetet werden, wo Text erlaubt ist. Er darf Zeichen enthalten, die ansonsten als Auszeichnungen oder Fluchtsymbol interpretiert werden. CDATA-Abschnitte beginnen mit der Zeichenfolge `<![CDATA[` und enden mit `]]>`.

Ein XML-Dokument sollte mit einer *XML-Deklaration* beginnen. Sie definiert die XML-Version und optionale Zeichensatzcodierung: `<?xml version="1.0" encoding=utf-8"?>` sind die Standardwerte, falls die Deklaration oder das encoding-Attribut fehlen.

Ein XML-Baum ist eine Baumrepräsentation eines XML-Dokuments mit einem Wurzelknoten, den *Dokumentknoten*. Neben der Einhaltung obiger Syntaxregeln, müssen XML-Dokumente immer eine korrekte Baumstruktur besitzen. Darüber hinaus darf nur ein Element – das Wurzelelement – sich unterhalb des Dokumentknotens befinden. Derartige XML Dokumente heißen *wohlgeformt*.

Schreiben von XML-Dokumenten Für die Auszeichnungen hat sich kein allgemein gültiger Stil etabliert. Am häufigsten werden Tags und Attribute klein geschrieben und mit dem Bindestrich getrennt, so wie in den Beispielen dieses Dokuments. Weniger häufig findet sich der uppper case camel style: `FirstName`.

XML-Dokumente sollen für den Menschen lesbar sein. Da die Elemente Daten auszeichnen, sollten möglichst aussagekräftige Substantive im Tag-Name verwendet werden. Abkürzungen vermeiden. Attribute sollten nur für Meta-Informationen verwendet werden. Innere Elemente einrücken.

Ergänzende Literatur Die Grundlagen für dieses Merkblatt finden sich in [1, Seiten 6–12] und für Details [2].

Literatur

[1] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management*. Cambridge University Press, New York, NY, USA, 2011. http://webdam.inria.fr/Jorge/files/wdm.pdf.

[2] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible markup language (XML) 1.0 (fifth edition). Technical report, W3C, November 2008. https://www.w3.org/TR/2008/REC-xml-20081126/.