

Markup-Sprachen und semi-strukturierte Daten, SS 2008
Übungsblatt 2

Abgabe: bis Do 01.05.2008, 14 Uhr

Besprechung am Fr 02.05.2008

Da der Donnerstag (01.05.) ein Feiertag ist, finden an diesem Tag keine Übungen statt. Beachten Sie aber, daß die Abgabe dennoch am Donnerstag zur üblichen Zeit fällig ist. Wir bitten Sie ferner, sich gleichmäßig auf die Übungen am Freitag (02.05.) zu verteilen.

Aufgabe 2-1 schriftlich bearbeiten
Document Object Model, XML Information Set

Machen Sie sich mit den W3C Empfehlungen zum Document Object Model (kurz: DOM) und zum XML Information Set vertraut.

Welche der folgenden Aussagen über das DOM, das XML Information Set und über XML Dokumente im Allgemeinen sind wahr?

- a) Das XML Information Set definiert eine abstrakte Menge an Daten. Es hat das primäre Ziel, eine konsistente Menge an Definitionen bereitzustellen, die in anderen Empfehlungen, welche Aussagen über die Informationen in einem XML Dokument treffen wollen, benutzt werden können.
- b) Das DOM ist eine sprachunabhängige Schnittstelle für Programme und Skripte. Es dient zum dynamischen Zugriff und zur Veränderung des Inhalts, der Struktur und des Stils von Dokumenten.
- c) Es gibt 4 aufeinander aufbauende und vom W3C genormte Versionen (Levels) des Document Object Models.
- d) Das DOM stellt Schnittstellen zu einer Menge von vordefinierten Objekten zur Repräsentation von HTML und XML Dokumenten bereit.
- e) Da DOM genau vorschreibt wie Datenstrukturen für XML Dateien implementiert werden, wird die Interoperabilität im Web unterstützt.
- f) Ein Unterschied zwischen dem DOM und dem XML Information Set ist, dass das XML Information Set keine Methoden zur Änderung des Inhalts eines XML Dokuments vorsieht.
- g) DOM Level 3 ist noch ein Working Draft.
- h) Mit "DOM Level 0" meint man die Funktionen zur Manipulation von HTML Dokumenten, die von Netscape Navigator 3.0 und Opera 1.0 bereitgestellt wurden.

- i) DOM Level 3 unterscheidet sich von DOM Level 2 unter anderem dadurch, dass eine komplette Abbildung zwischen dem XML Information Set und DOM bereitgestellt wird.
- j) DOM Level 3 besteht aus den vier Empfehlungen DOM Level 3 Core, DOM Level 3 HTML, DOM Level 3 Load and Save, DOM Level 3 Validation.
- k) Die Knotentypen in einer DOM-Repräsentation eines XML Dokuments (Dom Level 3) sind Document, DocumentFragment, DocumentType, EntityReference, ProcessingInstruction, Attr, Element, Comment, Text, CDataSection, Entity, Notation.
- l) Ein Document Node im DOM kann als Kinder maximal ein Element, beliebig viele Processing Instructions, beliebig viele Kommentare, beliebig viele Attribute und maximal eine Dokumenttypdeklaration haben.
- m) Ein Document Information Item im XML Information Set kann als Kinder maximal ein Element, beliebig viele Processing Instructions, beliebig viele Kommentare und maximal eine Dokumenttypdeklaration haben.
- n) CData Abschnitte in XML Dokumenten werden benutzt um Textblöcke, welche normal als XML interpretiert würden als reinen Text einzubinden. CData Abschnitte werden durch ']' beendet.
- o) Das Document Object Model sieht XML Attribute als keine tatsächlichen Knoten mit eigener Identität im XML Baum, sondern als Eigenschaften der Elemente, zu denen sie gehören. Deswegen sind die Eigenschaften parentNode, previousSibling, und nextSibling eines Attributobjekts im DOM mit null belegt. Das XML Information Set sieht Attribute dagegen als eigene Information Items an.
- p) Attributwerte eines XML-Elements erhält man im DOM Level 3 durch die Methode `getAttribute`, Attributknoten eines Elements durch die Methode `getAttributeNode`.

Aufgabe 2-2 schriftlich lösen

Umwandlung XHTML Tabelle in CSV

Comma-separated values (CSV)¹ ist ein häufig benutztes Dateiformat zum Austausch von tabellarischen Daten und wird z.B. von jeder Tabellenkalkulation unterstützt.

In XHTML werden Tabellen bekanntermaßen mit folgendem Markup geschrieben²:

```
<table>
  <tr> <th>Überschrift Spalte 1</th>
    <th>Überschrift Spalte 2</th>
    <!-- ... -->
  </tr>
  <tr> <td>Datenwert Spalte 1, Zeile 1</td>
    <td>Datenwert Spalte 2, Zeile 1</td>
    <!-- ... -->
  </tr>
  <tr> <td>Datenwert Spalte 1, Zeile 2</td>
    <td>Datenwert Spalte 2, Zeile 2</td>
    <!-- ... -->
  </tr>
```

¹siehe z.B. http://en.wikipedia.org/wiki/Comma-separated_values

²siehe z.B. <http://de.selfhtml.org/html/tabellen/index.htm>

```
<!-- ... -->
</table>
```

Schreiben Sie ein Java-Programm, das ein gültiges XHTML-Dokument als DOM-Baum einliest, darin die erste Tabelle findet und Sie im CSV-Format ausgibt. Die Ausgabe soll auf Standard-Out (`System.out`) erfolgen. Zwischen Überschriften und normalen Datenwerten wird nicht unterschieden. Zur Erleichterung darf angenommen werden, daß es sich um eine völlig reguläre Tabelle handelt, also z.B. keine Zellverbindungen (durch Attribut `colspan`) vorliegen und die Anzahl der Spalten in jeder Zeile gleich ist. Ferner darf angenommen werden, daß in den Einträgen der XHTML-Tabelle keine Kommas auftreten (d.h. Escaping nicht nötig) und auch kein weiteres Markup (wie z.B. `<p>`). Beachten Sie allerdings daß Einträge Zeilenumbrüche enthalten können, die sinnvoll behandelt werden müssen (z.B. ersetzen durch ein Leerzeichen).

Bei dem Übungsmaterial auf der Web-Seite zur Vorlesung findet sich ein Code-Gerüst für die Aufgabe, das sich um die Erzeugung eines DOM Baumes kümmert. Java ab Version 1.5 ist dafür nötig. Die benötigten Java-Klassen/-Interfaces zum Arbeiten mit DOM-Bäumen befinden sich im Paket `org.w3c.dom`³. Namespaces dürfen zur Vereinfachung in Ihrer Implementierung ignoriert werden. Ferner finden Sie beim Übungsmaterial XHTML-Eingabedokumente zum Testen. Wie immer gilt: Abgaben, die nicht einmal kompilieren, sind auch nichts wert.

Zusatzfrage (nicht schriftlich zu lösen): Vergleichen Sie CSV und XML/XHTML. Wo liegen die jeweiligen Vor- und Nachteile? Wie unterscheiden sich Ihre Einsatzbereiche?

Aufgabe 2-3 Semistrukturierte Daten

In dieser Aufgabe soll die Bezeichnung *Kurs* als Oberbegriff für *Vorlesung*, *Seminar* usw. stehen, die Bezeichnung *Dozent* für jegliche an einem Kurs beteiligte Lehrperson.

Dozenten der Vorlesung “Wissensrepräsentation und Inferenz für das Semantic Web” sind Prof. Dr. Hans Juergen Ohlbach und Michael Eckert. Dozenten für die Vorlesung “Logik für Informatiker” sind Prof. Dr. François Bry und Dr. Norbert Eisinger. Dozenten des Hauptseminars “Query Languages and Visualisation for the Semantic Web” sind Prof. Dr. François Bry und Dr. Norbert Eisinger und Tim Furche.

Repräsentieren Sie diese Informationen als semistrukturierte Daten. Die Repräsentation einer Person soll auch die Information enthalten, an welchen Kursen sie beteiligt ist. Überlegen Sie zunächst, für welche Teile der Information (explizite) Objektidentifikatoren sinnvoll sind. Überlegen Sie auch, welche Information geordnet bzw. ungeordnet repräsentiert werden soll.

Aufgabe 2-4 DOM und Javascript

In den Anlagen zum Übungsblatt finden sie eine HTML Datei `studentenleben.html`, in deren Body sich eine ungeordnete Liste befindet. Im Kopf der Datei befindet sich ein leeres `script` Element. Schreiben Sie in dieses Element ein Javascript Programm, welches mit Hilfe der DOM API die Liste zum Rotieren bringt. Zur Lösung dieser Aufgabe können Ihnen unter anderem die Methoden und Attribute `document.getElementById`, `removeChild`, `appendChild`, `window.setTimeout`, `nodeType` und `nextSibling` helfen.

```
<html>
<head>
  <title>Das Studentische Leben -- Ein Karusell der Abwechslung</title>
  <script type="text/javascript">
```

³Siehe Java-API Dokumentation <http://java.sun.com/javase/6/docs/api/>

```
</script>
</head>
<body>
  <h1>Das Studentische Leben</h1>
  <h3>Ein Karusell der Abwechslung</h3>
  <p>
    Der Student erfreut sich eines unbeschwerten und abwechslungsreichen Lebens:
  </p>
  <ul id="dieListe">
    <li><em>Studieren</em></li>
    <li>Essen</li>
    <li><strong>Feiern</strong></li>
    <li>Schlafen</li>
  </ul>
</body>
</html>
```