

html2pdf.xsl

XSL-FO für die Westentasche

für Facharbeiten, Hausaufgaben, Anleitungen, etc

Layman's L@tex

weil Wörd so kompliziert ist...

demonstriert und integriert die Open Source Tools

Apache FOP, Saxon und HTML Tidy

Motivation

Dieses Programm ist ein Trainingsprojekt für meinen neuen Job! Seit März 2008 bin ich Vollzeit XSLT Programmierer (ja, sowas gibt es ;-). Falls mich mal der Hafer sticht und ich nochmal eine Studienarbeit anfertigen sollte, werde ich mich nicht mehr Wörd oder Open Office herumärgern sondern ein XSL-FO Stylesheet programmieren, so wie dieses hier...



Inhaltsverzeichnis

Motivation.....	2
1 Technik.....	5
1.1 Systemanforderungen.....	5
1.2 Komponenten der Software.....	5
2 Formatierungsmöglichkeiten.....	7
2.1 Coverseite - <h1> Elemente.....	7
2.2 Textelemente.....	8
2.2.1 Ungeordnete Listen - Element.....	8
2.2.2 Hinweise - Element.....	9
2.2.3 Fussnoten - <i> Element.....	9
2.2.4 Tabellen - <table> Element.....	9
2.2.5 Vorformatierter Text - <pre> Element.....	11
2.2.6 Bilder - Element.....	11
2.2.7 Links - <a> Element.....	12
2.3 Kapitelstruktur - <h2>,<h3>,<h4>,<h5>.....	12
3 Lizenz für das Stylesheet html2pdf.xsl.....	13
Abbildungsverzeichnis.....	15
Literaturverzeichnis.....	17

1 Technik

Aus einem relativ unformatierten HTML Dokument wird mittels einer XSL Transformation ^[1] ein strukturiertes PDF Dokument generiert, mit Kapitelnummerierung, automatisch generierten Verzeichnissen für Inhalt, Abbildungen und Fussnoten. Weiterhin gibt es "running header" im Kopfbereich jeder Seite und eine Cover Seite. Das Format ist einerseits für einen zweiseitigen Druck ausgelegt (Die Abstände zum Seitenrand sind bei linker und rechter Seite verschieden und es gibt einen "Bundsteg"). Dieses Format wird mit book.bat erzeugt. andererseits gibt es aber auch die Möglichkeit ein PDF mit weniger Formatierungen zu generieren (paper.bat).

1.1 Systemanforderungen

Betriebssystem:

- WindowsXP / VISTA

Programmiersprachen:

- Java Runtime Environment



Selbstverständlich sollte das Programm auch unter einem anderen Betriebssystem laufen, nachdem die Komponenten der Software mit Java realisiert sind. Aber dann müssen die Start-Batchskripte angepasst werden. Hier übernehme ich keinen "Support" ;-)

1.2 Komponenten der Software

Es wird ausschliesslich Open Source Software ^[2] verwendet. Insbesondere die Tools:

- [Apache FOP](#)
- [Saxon](#)
- [Offo Hyphenation](#)
- [HTML Tidy](#)



Während im kommerziellen Umfeld Tools wie [Antenna House](#) eingesetzt werden, sind meiner Ansicht nach für einfachere Anwendungen, wie z.b. eine Diplomarbeit zu formatieren diese Tools vollkommen ausreichend. Gerne könnt ihr dieses Skript hier für solche Sachen verwenden.

Eigentlich ist das ganze ziemlich abgefahren, denn normalerweise werden die Formate HTML und PDF aus einer XML Quelle generiert. Hier ist es ein bisschen anders. Ein ganz einfaches HTML Dokument dient als Eingabe und das Programm generiert daraus über ein XML Zwischenformat das PDF Dokument. Das HTML Dokument kann per Hand oder mittels eines Editors (Ich verwende Dreamweaver 4) erstellt werden. Nachdem Saxon ein wohlgeformtes XML Dokument als Eingabe erwartet, aber gängige Editoren das selten erzeugen können, verwende ich noch HTML Tidy ^[3] um den HTML Code aufzuräumen - bevor er an Saxon übergeben wird.

[1] XSL Transformation, kurz XSLT, ist eine Programmiersprache zur Transformation von XML-Dokumenten. Sie ist Teil der Extensible Stylesheet Language (XSL) und stellt eine turing-vollständige Sprache dar... (mehr dazu auf [Wikipedia](#))

[2] Open source (engl.) bzw. quelloffen ist Software, die unter einer von der Open Source Initiative (OSI) anerkannten Lizenz steht. Die OSI stützt sich bei der Bewertung auf die Kriterien der Open Source Definition, die weit über die Verfügbarkeit des Quelltexts hinausgeht und fast deckungsgleich mit sog. Freier Software ist, d. h. der Quelltext muss auch offen für Bearbeitung und Weiterverbreitung sein... (mehr dazu auf [Wikipedia](#))

[3] HTML Tidy ist ein Computerprogramm, das dabei hilft, ungültige HTML-Tags zu erkennen und diese Fehler durch Entfernen bzw. Einfügen der fehlenden Tags zu beheben. (mehr dazu auf [Wikipedia](#))

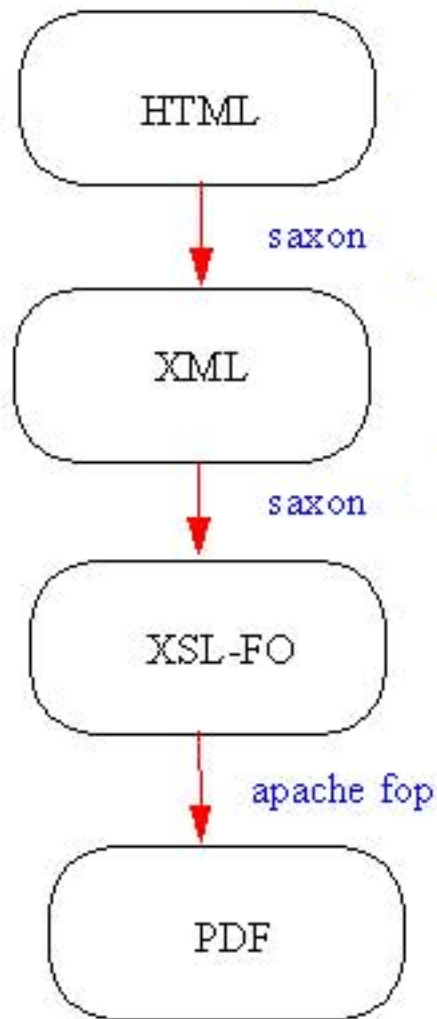


Abb. 1: Schema des dreistufigen Transformationsprozess

2 Formatierungsmöglichkeiten

Die folgenden HTML Tags werden verarbeitet:

- <h1> bis <h6>
- ,<table>,
- ,<i>,<a>
- <p>,<pre>



Das ganze ist mit Absicht so spartanisch gehalten, weil man dadurch das Dokument mit ganz einfachen Tools editieren kann, z.B. mit Notepad auf einem PDA oder mit dem Handy... Durch den Einsatz von HTML Tidy kann aber auch ein herkömmlicher Webeditor verwendet werden.

2.1 Coverseite - <h1> Elemente

Die Überschriften Heading 1 <h1> werden für die Texte auf der Coverseite verwendet:



Abb. 2: die Coverseite wird bedatet durch die ersten 7 Überschriften `<h1>`

2.2 Textelemente

2.2.1 Ungeordnete Listen - `` Element

Diese werden bis zu zwei Ebenen unterstützt, z.B.:

- item 1
- item 2
- item 3 mit Unterliste
 - item 1 der Unterliste
 - item 2 der Unterliste
 - item 3 der Unterliste
 - item 4 der Unterliste
 - item 5 der Unterliste
 - item 6 der Unterliste

Die HTML Eingabe würde für die Liste oben so aussehen:

Abb. 3: HTML Quellcode für eine Liste

```

1
2 <ul>
3   <li>item 1</li>
4   <li>item 2</li>
5   <li>item 3 mit Unterliste<br />
6     <ul>
7       <li>item 1 der Unterliste</li>
8       <li>item 2 der Unterliste</li>
9       <li>item 3 der Unterliste</li>
10      <li>item 4 der Unterliste</li>
11      <li>item 5 der Unterliste</li>
12      <li>item 6 der Unterliste</li>
13    </ul>
14  </li>
15 </ul>
16

```

2.2.2 Hinweise - Element

Hinweise werden mit dem Element gekennzeichnet - der folgende Absatz erscheint im PDF mit Hinweis Symbol und grau hinterlegt:



bla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla bla-
 bla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla bla-
 bla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla
 blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla bla

2.2.3 Fussnoten - <i> Element

Fussnoten werden mit dem <i> Tag gekennzeichnet. Das ist ein kleiner Missbrauch, denn <i> steht für italic und formatiert den Text normalerweise kursiv. Fussnoten funktionieren nur im normalen Fliesstext. Sie werden im Literaturverzeichnis nochmal aufgeführt, hier ist ein Absatz mit Fussnoten:

bla blaaa bla blabla blabla ^[4] blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla
 blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa bla blabla blabla blaaa
 bla blabla ^[5] lorem ipsum doloris lorem ipsum doloris lorem ipsum doloris lorem ipsum doloris ^[6]

2.2.4 Tabellen - <table> Element

Tabellen werden mit dem <table> Tag gekennzeichnet. Die PDF Transformation unterstützt einfache HTML Tabellen, einerseits zum Anordnen von Grafiken und Text für Layoutzwecke und andererseits im herkömmlichen Sinn.

Ohne die Angabe des border Attributes, bzw. mit einem Wert für das border Attribut von 0 erscheint die Tabelle ohne Rahmen und Hintergrund. Auf diese Art und Weise kann man Bilder anordnen, etc.

[4] bla bla wird gerne als Fülltext genommen

[5] zuviel bla bla stört aber

[6] da ist lorem ipsum schon besser


zelle 1	zelle 2	zelle 3	zelle 4
zelle mit liste:		bla	bla
<ul style="list-style-type: none"> - item 1 - item 2 - item 3 mit Unterliste <ul style="list-style-type: none"> - item 1 der Unterliste - item 2 der Unterliste 			
bla	bla	bla	bla

Abb. 4: HTML Quellcode für die Formatierung einer Tabelle ohne Rahmen

```

1
2 <table border="0">
3   <tbody>
4     <tr>
5       <td>zelle 1</td>
6       <td>zelle 2</td>
7       <td>zelle 3</td>
8       <td>zelle 4</td>
9     </tr>
10    <tr>
11      <td>
12        <p>zelle mit liste:</p>
13        <ul>
14          <li>item 1</li>
15          <li>item 2</li>
16          <li>item 3 mit Unterliste<br />
17            <ul>
18              <li>item 1 der Unterliste</li>
19              <li>item 2 der Unterliste</li>
20            </ul>
21          </li>
22        </ul>
23      </td>
24      <td>
25        
26      </td>
27      <td>bla</td>
28      <td>bla</td>
29    </tr>
30    <tr>
31      <td>bla</td>
32      <td>bla</td>
33      <td>bla</td>
34      <td>bla</td>
35    </tr>
36  </tbody>
37 </table>
38
39

```

Wenn die Tabelle mit Hintergrund und Rahmen im PDF erscheinen soll, dann muss man das border Attribut mit mindestens 1 setzen, hier dieselbe Tabelle zum Vergleich:

zelle 1	zelle 2	zelle	zelle 4
zelle mit liste: – item 1 – item 2 – item 3 mit Unterliste – item 1 der Unterliste – item 2 der Unterliste		bla	bla
bla	bla	bla	bla

Abb. 5: Gegenüberstellung von Werten

Der Untertitel der Tabelle, hier "Gegenüberstellung von Werten" wird durch das unmittelbar nach der Tabelle folgende `<p>` Tag angegeben.

2.2.5 Vorformatierter Text - `<pre>` Element

mit dem `<pre>` Element kann man Quelltext formatieren, so wie diesen hier:

Abb. 6: Hello World Programm

```

1
2 <pre>
3 class HelloWorld {
4     public static void main(String[] args) {
5         System.out.println("Hello World!");
6     }
7 }
8 </pre>
9 <p>Hello World Programm</p>
10

```

Der Absatz `<p>` der unmittelbar nach dem `<pre>` Element kommt, wird vom PDF Stylesheet als Titel interpretiert

2.2.6 Bilder - `` Element

Bilder werden mit dem `` Tag eingebunden. Übrigens sind die kleinen icons von der Website simplebits.com. Ich habe eine Lizenz für sie erworben. Diese dürfen nicht weiterverwertet werden.



Abb. 7: Ein Diskette Icon - die Dinger gibt es schon nicht mehr...



Abb. 8: Screenshot vom Handy-Mandarin-Trainer

Ein Absatz der dem Bild folgt wird als Untertitel interpretiert - das gilt nicht für Bilder in Tabellen.



Grafiken werden am besten im Verzeichnis 'images' abgelegt. Das Logo und die icon Grafiken von book.pdf befinden sich im Ordner 'stylesheet/styleimages'.

2.2.7 Links - <a> Element

Hyperlinks wie dieser [hier zu meiner Homepage](#), erscheinen im PDF wie im HTML gleich...

2.3 Kapitelstruktur - <h2>,<h3>,<h4>,<h5>

Die Kapitelstruktur wird aus den Überschriften generiert. Dabei werden die Überschriften automatisch nummeriert! Es werden nur die Überschriften Heading 2 <h2> bis Heading 5 <h5> verarbeitet.

Ein bisschen muss man dabei aufpassen:



Auf eine Überschrift Heading 2 <h2> muss eine Überschrift heading 3 <h3> folgen, auf <h3> kann <h4> oder <h2> folgen, auf <h4> kann <h4>, <h3> oder <h2> folgen, usw. Ausserdem sollte das Dokument mit einer Überschrift Heading <h2> beginnen... (nach den <h1> für die Coverseite).
ACHTUNG: Das Programm kann kein PDF erzeugen, wenn man sich hier vertut!

3 Lizenz für das Stylesheet html2pdf.xsl

THIS SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.



Abbildungsverzeichnis

1	Schema des dreistufigen Transformationsprozess.....	6
2	die Coverseite wird bedatet durch die ersten 7 Überschriften <h1>.....	8
3	HTML Quellcode für eine Liste.....	9
4	HTML Quellcode für die Formatierung einer Tabelle ohne Rahmen Gegenüberstellung von Werten.....	10
5	Gegenüberstellung von Werten.....	11
6	Hello World Programm.....	11
7	Ein Diskette Icon - die Dinger gibt es schon nicht mehr.....	11
8	Screenshot vom Handy-Mandarin-Trainer.....	12



Literaturverzeichnis

1	XSL Transformation, kurz XSLT, ist eine Programmiersprache zur Transformation von XML-Dokumenten. Sie ist Teil der Extensible Stylesheet Language (XSL) und stellt eine turing-vollständige Sprache dar... (mehr dazu auf Wikipedia).....	5
2	Open source (engl.) bzw. quelloffen ist Software, die unter einer von der Open Source Initiative (OSI) anerkannten Lizenz steht. Die OSI stützt sich bei der Bewertung auf die Kriterien der Open Source Definition, die weit über die Verfügbarkeit des Quelltexts hinausgeht und fast deckungsgleich mit sog. Freier Software ist, d. h. der Quelltext muss auch offen für Bearbeitung und Weiterverbreitung sein... (mehr dazu auf Wikipedia).....	5
3	HTML Tidy ist ein Computerprogramm, das dabei hilft, ungültige HTML-Tags zu erkennen und diese Fehler durch Entfernen bzw. Einfügen der fehlenden Tags zu beheben. (mehr dazu auf Wikipedia).....	5
4	bla bla wird gerne als Fülltext genommen.....	9
5	zuviel bla bla stört aber.....	9
6	da ist lorem ipsum schon besser.....	9