Kurzanleitung für das Konvertieren von XML Dokumenten in PDF Dokumente

Ylva Brandt, <y.brandt@tu-bs.de>
TU Braunschweig

14. Juni 2004

Übersicht

Im Rahmen des Portiko Projektes wurde an der TU-Braunschweig bereits ein Konvertierungswerkzeug entwickelt, das es ermöglicht XML Dokumente in HTML Dokumente umzuwandeln. Dieses Konvertierungswerkzeug erlaubt nun auch die Umwandlung der XML Dokumente in das Portable Document Format (PDF), um es in druckbarer Form bereitzustellen.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Allgemeines

1.1 Beschreibung

Das vorliegende Dokument erläutert die Umwandlung eines XML Dokumentes, das gemäß der contentobject – DTD erstellt wurde, in das Portable Document Format (PDF). Das PDF ermöglicht die Veröffentlichung eines Kurses in druckbarer Form.

Die Umwandlung in das PDF erfolgt über ein Latex Dokument, das zunächst aus dem XML Dokument erzeugt wird. Aus dem Latex Dokument wird anschließend mit Hilfe von *pdflatex* das PDF Dokument erzeugt. Diese Konvertierungsvorgänge werden wie auch die Umwandlung nach HTML mit dem im Rahmen der XML-AG entwickelten *Content Converter* durchgeführt. Dabei sind jedoch einige zusätzliche Hinweise zu beachten, damit sowohl die Umwandlung nach HTML als auch ins PDF ein zufriedenstellendes Ergebnis liefern. Das vorliegende Dokument stellt daher eine Ergänzung zur „Kurzanleitung für das Publizieren von XML Dokumenten auf der Hyperwave Plattform“ [1] dar.
Die Umwandlung ins PDF erfolgt jedoch unabhängig von der Umwandlung nach HTML, so dass es dem Anwender freigestellt ist, welche Ausgabeformate er erzeugen lässt.

Um Fehlerquellen bei der Umwandlung der Dokumente zu vermeiden, wird empfohlen, sowohl diese als auch die obengenannte Anleitung vollständig durchzulesen und vor allem die Hinweise und bekannten Fehler zu berücksichtigen.

1.2 Benötigte Programme

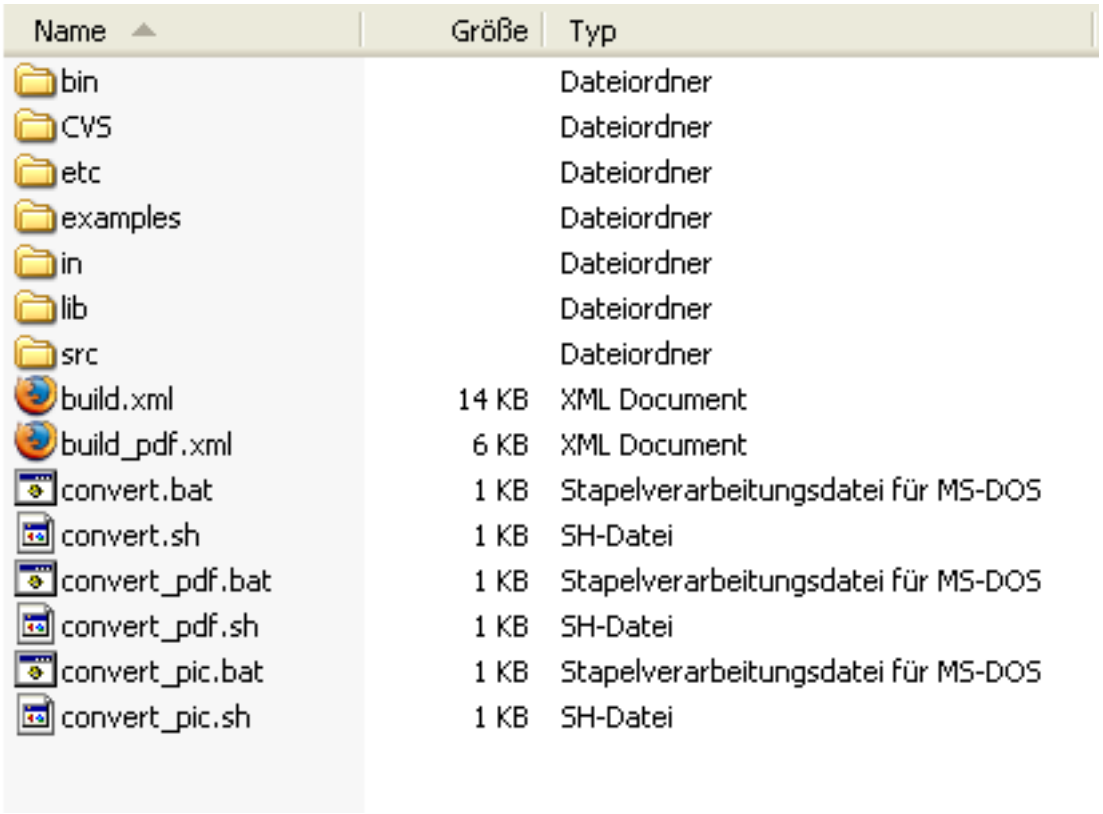
1.2.1 Content Converter

Der *Content Converter* ist gemäß den Anweisungen in [1] zu installieren.

An der grundsätzlichen Verzeichnisstruktur des *Content Converters* ändert sich für die Konvertierung ins PDF nichts. Lediglich die für die Umwandlung relevanten Dateien sind andere.

Die einzelnen Unterverzeichnisse haben folgende Aufgaben:

- *bin:* Das Startskript „build_pdf.bat“ bzw. „build_pdf.bat“ starten die Umwandlung. Diese Dateien sind jedoch für den Benutzer uninteressant.
- *etc:* Hier lagern neben den XSLT und CSS Stylesheets für die Umwandlung nach HTML auch die XSLT Stylesheets für die Umwandlung nach Latex. In dem Unterordner „stylesheets/ltx_styles“ liegen Stildateien, die von Latex für die Erstellung der Verzeichnisse benutzt werden. Diese Datein sollten jedoch vom Anwender nicht verändert werden.
- *in:* In dieses Verzeichnis wird das umzuwandelnde Dokument kopiert.



Name	Größe	Typ
bin		Dateiordner
CVS		Dateiordner
etc		Dateiordner
examples		Dateiordner
in		Dateiordner
lib		Dateiordner
src		Dateiordner
build.xml	14 KB	XML Document
build_pdf.xml	6 KB	XML Document
convert.bat	1 KB	Stapelverarbeitungsdatei für MS-DOS
convert.sh	1 KB	SH-Datei
convert_pdf.bat	1 KB	Stapelverarbeitungsdatei für MS-DOS
convert_pdf.sh	1 KB	SH-Datei
convert_pic.bat	1 KB	Stapelverarbeitungsdatei für MS-DOS
convert_pic.sh	1 KB	SH-Datei

Abbildung 1.1: Hauptverzeichnis der Installation

- *lib:* In diesem Verzeichnis befinden sich weitere zur Umwandlung benötigte Programme. Für die Konvertierung ins PDF ist lediglich das Programm *ConvertImg* hinzugekommen, welches das Einbinden von Graphiken ermöglicht. Anpassungen sind hier in der Regel nicht erforderlich.
- *out:* Bei der Umwandlung ins PDF wird hier am Ende lediglich die Datei „output.pdf“ abgelegt, die das gesamte Dokument in druckbarer Form enthält.
- *src:* Hier liegt der Quellcode für das Programm *ConvertImg*. Der Quellcode ist aber für den Umwandlungsprozess irrelevant.

1.2.2 pdfLatex

Damit die vom *Content Converter* erzeugte Latex-Datei weiter ins PDF umgewandelt werden kann, muss pdfLatex auf dem Rechner installiert sein. Für das Betriebssystem „Windows“, kann Miktex verwendet werden.
Folgende Pakete werden in dem Latex Dokument benutzt und müssen daher zur Verfügung stehen. In Klammern sind jeweils die benötigten Dateien angegeben.

- KOMA-Script (scrartcl.cls, scrlfile.sty, typearea.sty, scrpage2.sty)
- ngerman (ngerman.sty)
- float (float.sty)
- graphicx (graphicx.sty, keyval.sty, graphics.sty, trig.sty, graphics.cfg, pdftex.def)
- ucs (ucs.sty, uniglobal.def)
- inputenc (inputenc.def, utf8.def)
- makeidx (makeidx.sty)
- multind (multind.sty)
- array (array.sty)
- tabularx (tabularx.sty)
- dcolumn (dcolumn.sty)
- enumerate (enumerate.sty)
- longtable (longtable.sty)
- hyperref (hyperref.sty, pd1enc.sty, hyperref.cfg, url.sty, hpdftex.def, pifont.sty, upzd.fd, upsy.fd)
- listings (listings.sty, lstpatch.sty, lstmisc.sty, listings.cfg)

Bei Miktex in der Version 2.4 oder höher sind diese Pakete standardmäßig enthalten.

1.2.3 ImageMagick

ImageMagick ist ein Programm zur Bildbearbeitung, das Kommandozeilen orientiert arbeitet. Dieses Programm wird benötigt, um alle Graphik-Dateien vor dem einbinden ins PDF zu konvertieren, da Latex die Orignaldateien nicht ohne weiteres verarbeiten kann. Sofern das Programm ImageMagick installiert wurde und die nötigen Umgebungsvariablen gesetzt wurden, wird die Konvertierung automatisch vom *Content Converter* durchgeführt.
ImageMagick ist ein frei verfügbares Programm, das von der Internetseite www.imagemagick.org bezogen werden kann. Dort sind auch die Installationsanweisungen und zu setzenden Umgebungsvariablen zu finden.

2 Umwandlung von XML Dateien in PDF Dokumente

2.1 Vorarbeiten

Die Umwandlung in ein PDF Dokument erfordert die gleichen Vorarbeiten wie die Umwandlung nach HTML, die in [1] beschrieben sind. Soll keine zusätzliche Konvertierung nach HTML stattfinden, kann auf das Kopieren der Bilder und Flash Animationen in das Verzeichnis *out* verzichtet werden. Die Bilder werden direkt in das PDF Dokument eingebunden und Flash Animationen können in einer Druckversion des Dokumentes verständlicherweise nicht angezeigt werden.

2.2 Start der Umwandlung

Die Umwandlung von XML nach PDF erfolgt in mehreren Schritten, die nacheinander durchgeführt werden müssen. Damit der Anwender nicht jeden Schritt einzeln ausführen muss, liegt im Hauptverzeichnis des *Content Converters* eine XML Datei, die den Ablauf steuert. Gestartet wird die Umwandlung mit Hilfe eines Skriptes, das ebenfalls im Hauptverzeichnis liegt. Dieses Skript erwartet als Parameter den Namen des zu konvertierenden XML Dokumetes (ohne den Verzeichnisnamen „in“).

Um das Skript aufzurufen, muss der Anwender zunächst eine Eingabeaufforderung starten und in das Hauptverzeichnis des *Content Converters* wechseln. Von dort aus kann nun das Skript `convert_pdf` gestartet werden. Soll zum Beispiel die Datei `beispiel.xml` ins PDF konvertiert werden, sieht der Aufruf wie folgt aus:

```
convert_pdf beispiel.xml
```

2.3 Ergebnis der Umwandlung

Das Ergebnis der Umwandlung ist die Datei *output.pdf*, die nach der Konvertierung im Verzeichnis *out* zu finden ist. Der Name der Ausgabedatei ist bei der Konvertierung auf „output“ festgelegt. Die Datei kann jedoch nach der Konvertierung beliebig umbenannt werden.

3 Hinweise zur Verwendung einzelner Elemente

3.1 Links und Querverweise

Es können zwar alle Arten von Links, die in [1] beschrieben sind, verwendet werden, ohne dass die Konvertierung ins PDF scheitert. Die internen Links auf Bilder, etc. mit dem Element `<link>` werden jedoch nicht unterstützt, da diese wie externe Links behandelt werden müssten. Das heißt, auch Bilder würden in einem Browser geöffnet. Dieser findet die entsprechende Datei jedoch nicht, wenn nur relative Pfadangaben vorhanden sind. Daher wird bei Verweisen auf Bilder mit dem Element `<link>` nur der enthaltene Text ausgegeben. Interne Verweise, die auch in der PDF Version des Dokumentes vollständig angezeigt werden sollen, sollten daher mit dem Element `<crossref>` erstellt werden.

Das Attribut `color` im Element `<link>` wird nicht unterstützt, da in der Latex Datei die Linkfarbe nur global festgelegt werden kann. Hyperlinks sind daher auf cyan und interne Verweise auf blau festgelegt.

Interaktive Verweise, die der Anwender beim Betrachten des PDF Dokumentes am Bildschirm anklicken kann, basieren auf den Nummern der einzelnen Kapitel, Abschnitte, etc. Wenn interaktive Verweise im PDF gewünscht werden, müssen daher das Attribut `numbering` im Element `<contentobject>` und das Attribut `number` im Element `<crossref>` den Wert `yes` haben. Ansonsten werden alle erwünschten Verweisdaten ebenfalls korrekt angezeigt, aber ein interaktiver Sprung zum Verweisziel ist in diesem Fall nicht möglich.

3.2 Bilder

3.2.1 Unterstützte Formate

Alle Verzeichnisse im Ordner *in* werden nach Bildern durchsucht, um sie in PDF Dateien zu konvertieren. Die Bilder werden an den Dateiendungen erkannt. Standardmäßig werden folgende Dateitypen unterstützt:
png, gif, jpg, tif, tiff, bmp, pcx, ps, eps und pdf
Sollten weitere Formate benötigt werden, müssen diese in der Datei „build_pdf.xml“ im

Target *convert_ img2pdf* beim Aufruf des Programms *ConvertImg* als Parameter ergänzt werden.

Da *pdfLatex* nicht in der Lage ist, Verzeichnispfade mit Punkten im Verzeichnisnamen zu interpretieren, sollte bei der Benennung von Ordnern auf den Punkt verzichtet werden. Nicht zulässig wäre also zum Beispiel:

```
<figure fileref='bilder/fotos_03.05.04/party.jpg' ...>
```

Stattdessen könnte man zum Beispiel folgende Benennung wählen:

```
<figure fileref='bilder/fotos_03_05_04/party.jpg' ...>
```

3.2.2 Größenangaben

Wenn das Dokument ins PDF konvertiert werden soll, ist zu beachten, dass eine DIN A4 Seite anders als der Browser nur beschränkten Platz bietet. Da *pdfLatex* nicht in der Lage ist, die Größe eines Bildes selbständig zu erfassen und gegebenenfalls zu skalieren, werden Bilder, die größer sind als eine DIN A4 Seite, einfach abgeschnitten. Um dies zu verhindern, sollten die Graphiken entweder von vorneherein klein genug sein oder es sollte mit Hilfe der Attribute `width` und / oder `height` des `<figure>` Elementes eine sinnvolle Größe angegeben werden. Unter Berücksichtigung der Seitenränder ergeben sich eine maximale Breite von 14,5 cm und eine maximale Höhe von 21,5 cm.

3.3 Tabellen

Bei der Erstellung von Tabellen sind ebenfalls einige Einschränkungen zu beachten. Zunächst einmal sollte die Verwendung der Attribute `colspan` und `rowspan` weitgehend vermieden werden, da sie von den Stylesheets zur Konvertierung ins PDF nicht unterstützt werden (abgesehen vom Attribut `colspan` im Tabellenkopf). Die Verwendung führt zwar nicht zum Abbruch der Konvertierung, die Zellen der Tabelle werden jedoch höchst wahrscheinlich verschoben dargestellt.

Wie bei den Bildern ist auch bei den Tabellen zu beachten, dass diese auf einer DIN A4 Seite nicht beliebig breit sein können. Am sichersten ist es daher, die Breite der einzelnen Spalten oder zumindest der gesamten Tabelle festzulegen. Die Angabe der Größe in Prozent bietet hier die Möglichkeit einer flexiblen Darstellung, je nach vorhandenem Platz.
Wird die Größe der Tabelle mit den Elementen `<colgroup>` und `<col>` definiert, werden diese Angaben verwendet und weitere Angaben in den Elementen `<table>`, `<th>` und `<td>` ignoriert. Ist keine Definition der Tabellengröße durch die Elemente `<colgroup>` und `<col>` vorhanden, so werden die Angaben in den Elementen `<table>`, `<th>` und `<td>`, falls vorhanden, wie folgt interpretiert.
Der Wert des Attributs `width` im Element `<table>` bestimmt die Gesamtbreite der Tabelle. Sollte dieser Wert außerhalb des darstellbaren Bereichs liegen, wird die maximale

Breite von 14,5 cm verwendet. Da in Latex die Breite und Ausrichtung einer Spalte insgesamt und nicht für jede einzelne Zelle definiert werden muss, wird die Breiten- und Ausrichtungsangabe der Zelle in der ersten vollbesetzten Zeile des Tabellenköpers repräsentativ für die gesamte Spalte betrachtet. Die so festgelegte Ausrichtung des Textes kann allerdings für einzelne Zellen später überschrieben werden. Sollte lediglich die Gesamtbreite der Tabelle angegegeben sein, wird diese gleichmäßig auf alle Spalten aufgeteilt. Sollte es sich bei der Spaltenbreite um absolute Angaben handeln, werden eventuelle Angaben zur Gesamtbreite der Tabelle ignoriert.
Fehlen hingegen sämtliche Breitenangaben, so wird die Breite der einzelnen Spalten automatisch von Latex festgelegt, wobei der Inhalt einer Zelle jedoch weder automatisch noch durch Angabe des Elementes `<br/>` umgebrochen werden kann, so dass bei langen Einträgen die Tabelle nicht vollständig dargestellt wird.

3.4 Multimedia

Da multimediale Inhalte, wie Ton, Video oder Animationen nicht gedruckt werden können, werden diese nicht direkt in das PDF Dokument eingebunden. Stattdessen wird nur die Beschreibung, falls vorhanden, angezeigt, und auf die entsprechende Multimedia Datei verwiesen. Ist eine *Caption* vorhanden wird der Name der Datei in Klammern angegeben, ansonsten dient der Name der Datei als *Caption*.

Literatur

[1] C. Klinzmann, „Kurzanleitung für das Publizieren von XML Dokumenten auf der Hyperwave Plattform“,

http://www.ibr.cs.tu-bs.de/arbeiten/cwerner/portiko-latex/kurzanleitung.pdf