

Dokumentautomation mit XML am Beispiel einer Banddiskografie

Hochschulinformationstag in Merseburg – 13. April 2013

Dr. Thomas Meinike

Hochschule Merseburg | FB Informatik und Kommunikationssysteme

<http://www.iks.hs-merseburg.de/~meinike/>

thomas.meinike@hs-merseburg.de

Zur Person

- Lehrkraft für besondere Aufgaben seit 1997
- Tätig in den Studiengängen im Fachbereich IKS:
 - Bachelor Technische Redaktion und E-Learning-Systeme
 - Master Technische Redaktion und Wissenskommunikation
- Lehr- und Arbeitsgebiete:
 - Auszeichnungs- und Skriptsprachen
 - Online-Dokumentation und Web-Entwicklung
 - **XML-Technologien**



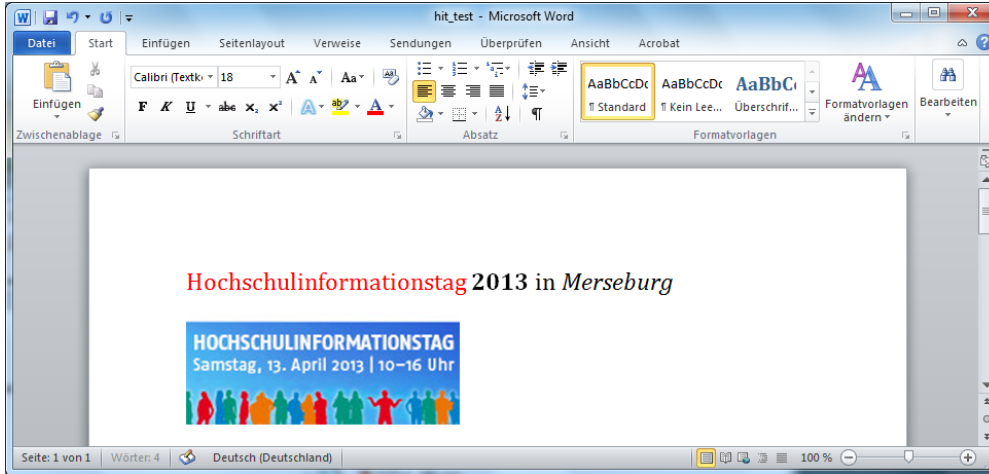
Technische Redaktion in Merseburg

- Diplomstudiengang 1995 bis 2008
Kommunikation und Technische Dokumentation
- Masterstudiengang seit 2006
Technische Redaktion und Wissenskommunikation (4 Sem. / M. A.)
- Bachelorstudiengang seit 2010
Technische Redaktion und
E-Learning-Systeme
(6 Sem. / B. Eng.)
- Weitere Details zu den
Inhalten gern im Anschluss.

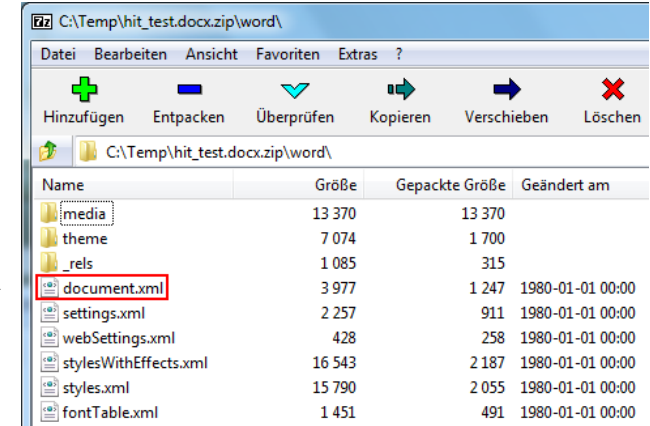


Praktischer Einstieg (»In Word steckt XML drin!«)

→ Word-Dokument mit Text, Formatierungen und Bild erstellen:



→ Dateiendung *.docx* in *.zip* umbenennen und in 7-Zip öffnen (*document.xml* enthält den obigen Text):



Praktischer Einstieg (»In Word steckt XML drin!«)

→ Lässt sich ein Word-Dokument aus den Bestandteilen im Archiv auch **ohne Word** erzeugen?



The screenshot shows an 'Archiv-Browser' window with a file tree on the left and an XML editor on the right. The file tree shows a folder 'hit_test.docx' containing several subfolders and files, with 'document.xml' highlighted in a red box. The XML editor shows the content of 'document.xml' with line numbers 28 to 53. The XML code is as follows:

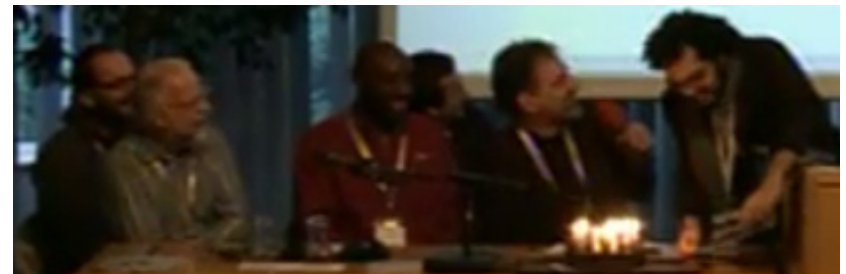
```
28 <w:r w:rsidRPr="00C51FF7">
29   <w:rPr>
30     <w:rFonts w:asciiTheme="majorHAnsi" w:hAnsiTheme="majorHAnsi"/>
31     <w:color w:val="FF0000"/>
32     <w:sz w:val="36"/>
33     <w:szCs w:val="36"/>
34   </w:rPr>
35   <w:t>Hochschulinformationstag</w:t>
36 </w:r>
37 <w:r w:rsidRPr="00C51FF7">
38   <w:rPr>
39     <w:rFonts w:asciiTheme="majorHAnsi" w:hAnsiTheme="majorHAnsi"/>
40     <w:sz w:val="36"/>
41     <w:szCs w:val="36"/>
42   </w:rPr>
43   <w:t xml:space="preserve"> </w:t>
44 </w:r>
45 <w:r w:rsidRPr="00C51FF7">
46   <w:rPr>
47     <w:rFonts w:asciiTheme="majorHAnsi" w:hAnsiTheme="majorHAnsi"/>
48     <w:b/>
49     <w:sz w:val="36"/>
50     <w:szCs w:val="36"/>
51   </w:rPr>
52   <w:t>2013</w:t>
53 </w:r>
```

Ja!

XML ...

- **Ex**ensible **M**arkup **L**anguage
- ... wurde 1998 vom World Wide Web Consortium (W3C) eingeführt
- ... ist keine eigenständige Sprache, sondern ein Konzept zur Definition von konkreten Auszeichnungssprachen
- ... definiert (relativ wenige) Regeln, welche beim Schreiben und der anschließenden Verarbeitung exakt einzuhalten sind.

15 Jahre XML (Prag, 10.02.13)



XML ...

→ Grundaufbau der Banddiskografie:

```
<?xml version="1.0" encoding="UTF-8"?>
<diskografie bandname="...">
  <bandinfo>Text</bandinfo>
  <referenzen> <referenz url="http://...">Linktext</referenz> </referenzen>
  <werk jahr="..." typ="...">
    <werkname>Text</werkname>
    <kommentar>Text</kommentar>
    <coverbild breite="..." hoehe="...">bildname.jpg</coverbild>
    <titelliste>
      <titel spielzeit="...">Text</titel>
    </titelliste>
  </werk>
</diskografie>
```

In Kürze:

Die Dokument-Struktur entsteht durch (verschachtelte) **Elemente**, welche mittels Tags markiert werden.

Dazwischen steht der eigentliche Inhalt.

`<elname>...</elname>`

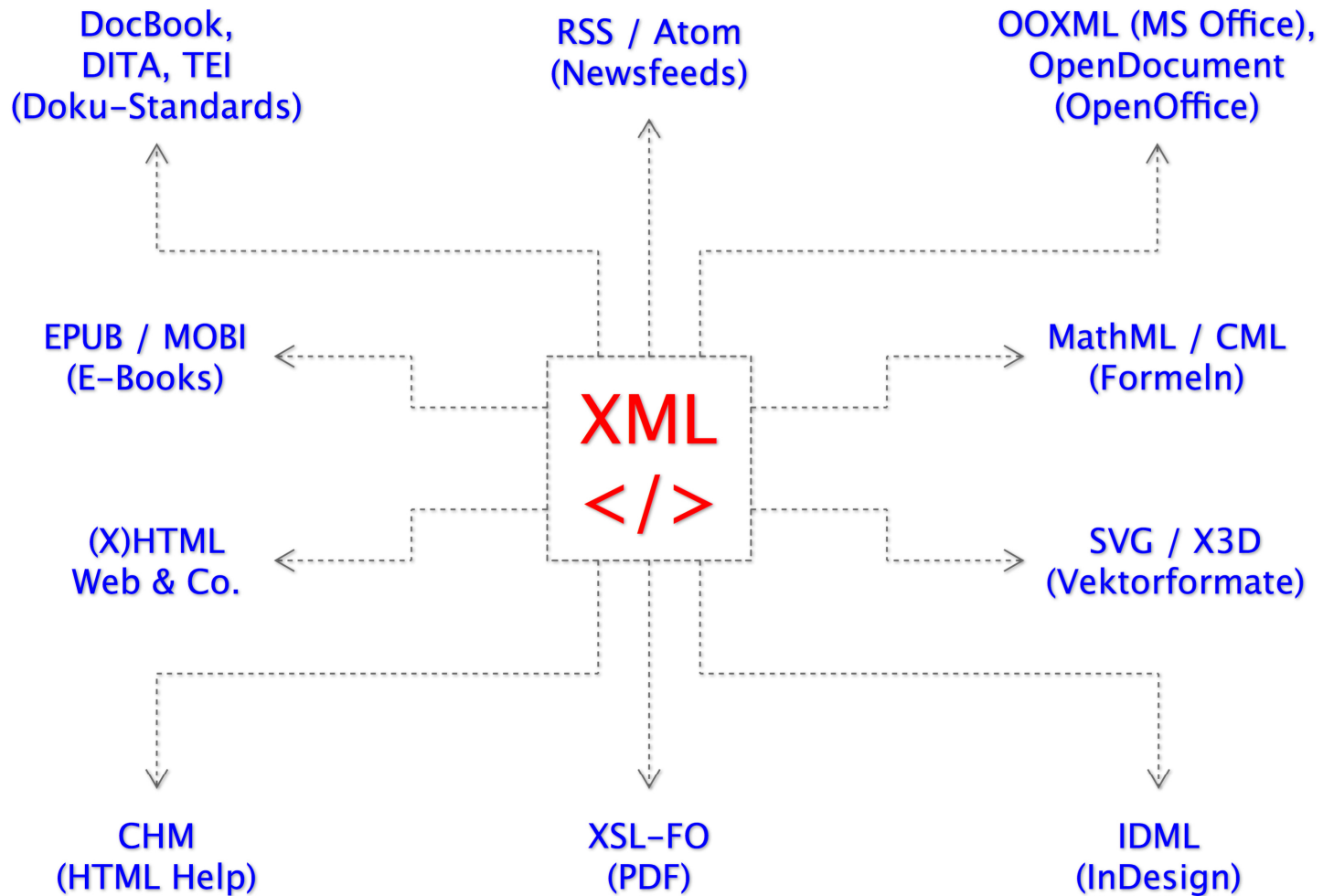
Attribute in der Form `atname="atwert"` enthalten zusätzliche Informationen.

Spezielle Datenmodelle helfen beim Bearbeiten und Prüfen.

XML ...

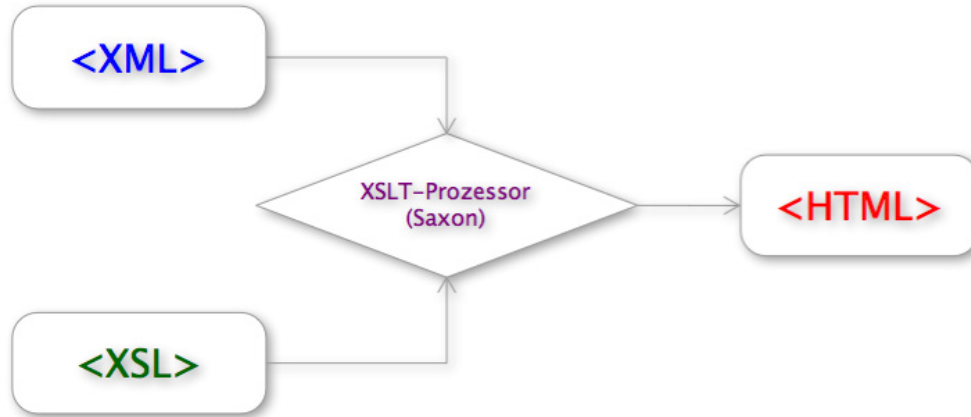
- ➔ ... bildet mit weiteren Konzepten wie XSLT, XSL-FO und XPath eine Technologiefamilie
- ➔ ... wird mittlerweile für vielfältige Anwendungen, speziell in der Technischen Dokumentation, eingesetzt (Single-Source-Publishing)
- ➔ ... unterstützt eigene Entwicklungen neben der Anwendung »gebrauchsfertiger« Sprachen (wie HTML)
- ➔ ... ermöglicht in Bereichen wie Technische Redaktion und Verlagswesen automatisierte Umsetzungen von Webinhalten, Onlinehilfen, PDF-Dateien und E-Books.

XML-Anwendungen (→ TR-Studium)



XML-Verarbeitung

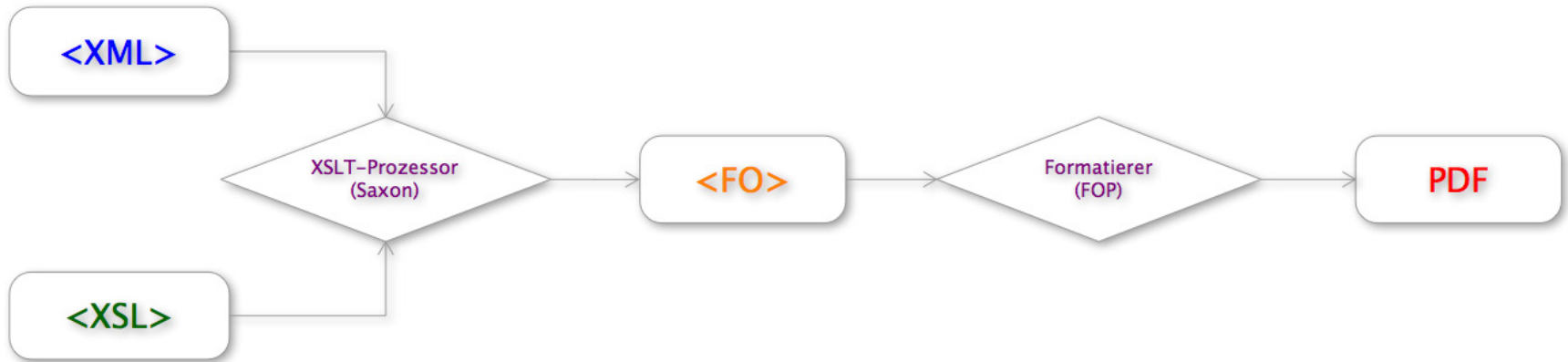
→ **HTML**-Generierung nach diesem Ablaufschema:



- Die XML-Struktur wird mit einem XSL(T)-Stylesheet (= Programm) in das Ausgabeformat HTML transformiert.
- HTML ist in Medienprodukten wie EPUB und CHM enthalten.

XML-Verarbeitung

→ Prozesskette der automatisierten **PDF**-Produktion:



- Hier wird zunächst ein Zwischenformat (XSL-FO) erzeugt und dieses im zweiten Schritt zum fertigen PDF verarbeitet.

XML-Verarbeitung

➔ Produktion von Archivformaten wie **DOCX, EPUB, IDML, CHM**:

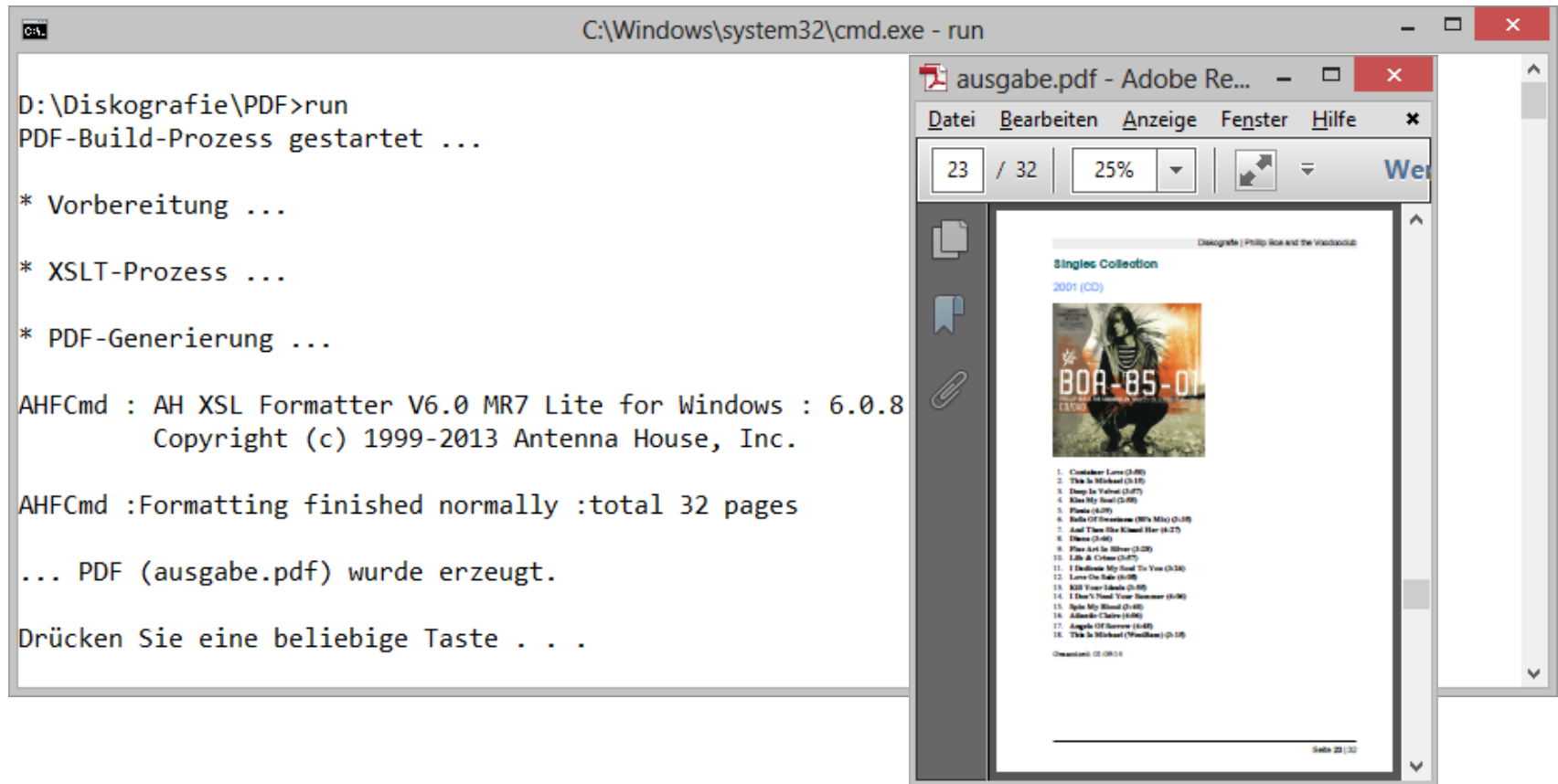
- Detaillierte Analyse der Dokumentstrukturen
- Anlegen der Zielstruktur
- Generierung der Inhalts- und Zusatzdateien (XSLT)
- Hinzufügen statischer Dateien, z. B. Bilder und Fonts
- Kompression mit ZIP-Werkzeug
- Dateiendung .zip umbenennen (→ .docx, .epub, .idml)

- Für das binäre Hilfeformat CHM (HTML Help) wird zusätzlich ein externer Hilfecompiler benötigt (→ .chm).

- Der automatisierte Ablauf wird mittels Batch- bzw. Shellskript gesteuert (siehe Demonstrationen).

XML-Verarbeitung

➔ Beispiel PDF-Produktion:



The image shows a Windows command prompt window and an Adobe Reader window. The command prompt window displays the following text:


```
C:\Windows\system32\cmd.exe - run
D:\Diskografie\PDF>run
PDF-Build-Prozess gestartet ...
* Vorbereitung ...
* XSLT-Prozess ...
* PDF-Generierung ...
AHFCmd : AH XSL Formatter V6.0 MR7 Lite for Windows : 6.0.8
          Copyright (c) 1999-2013 Antenna House, Inc.
AHFCmd :Formatting finished normally :total 32 pages
... PDF (ausgabe.pdf) wurde erzeugt.
Drücken Sie eine beliebige Taste . . .
```

The Adobe Reader window shows the PDF content, which is a discography for Philip Sosa and the Woodstock. The content includes a title "Single Collection", a subtitle "2001 (CD)", a cover image, and a list of 18 tracks with their durations.

Discografie | Philip Sosa and the Woodstock

Single Collection

2001 (CD)

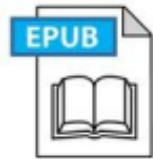


1. Consider Love (3:40)
2. Think In Michael (3:17)
3. Deep In Value (3:07)
4. Kiss My Soul (3:00)
5. Please (3:07)
6. Back Of Tomorrow (8th Mix) (3:18)
7. And Then She Kissed Her (4:27)
8. Please (3:40)
9. Please And In Silver (3:20)
10. Life & Love (3:07)
11. I Believe My Soul To You (3:24)
12. Love On Side (3:08)
13. Kiss Your Soul (3:40)
14. I Don't Need Your Summer (3:46)
15. Split My Soul (3:40)
16. Adrenaline Choke (4:06)
17. Angels Of Heaven (4:40)
18. Think In Michael (Woodman) (3:18)

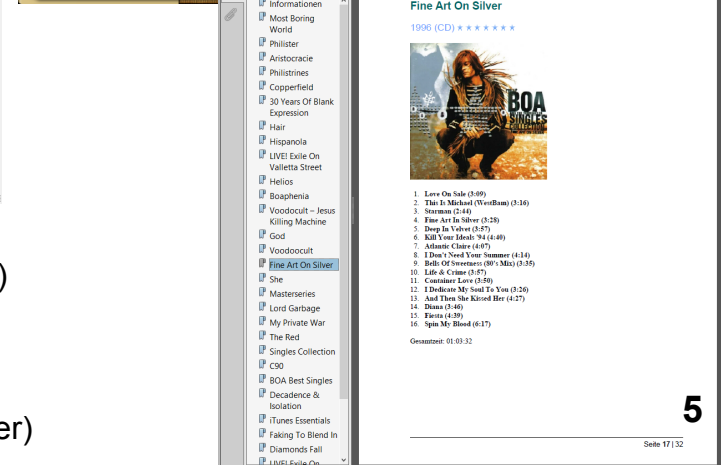
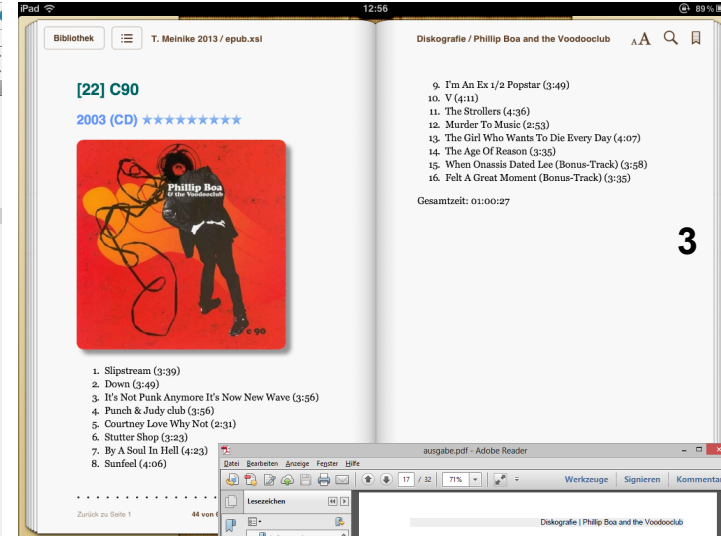
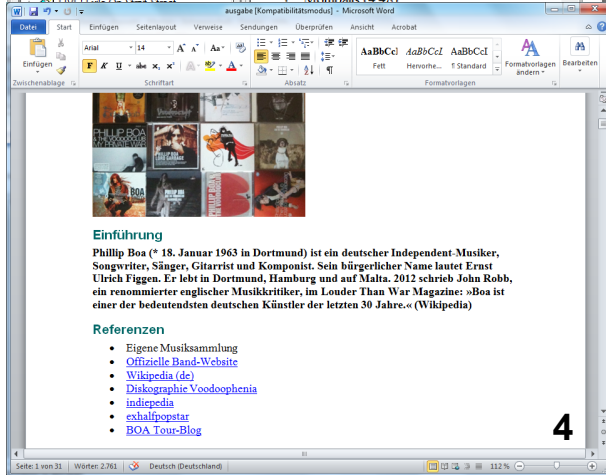
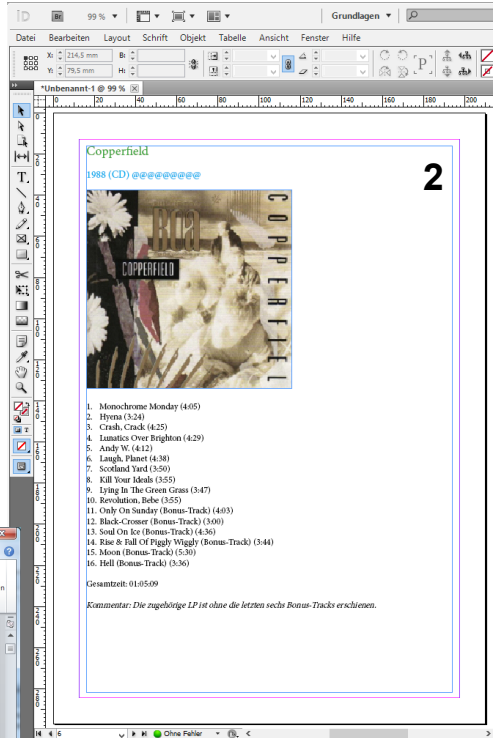
Downloaded: 02.09.14

Seite 23 | 32

{Praktische Demonstration ...}



Diskografie-Ausgaben



- 1 = CHM (Help Viewer)
- 2 = IDML (InDesign)
- 3 = EPUB (iBooks)
- 4 = DOCX (Word)
- 5 = PDF (Adobe Reader)

Zusammenfassung und Ausblick

- ➔ XML und zugehörige Konzepte haben sich in der Informations- und Wissensgesellschaft fest etabliert.
- ➔ Das Anwendungsspektrum reicht von der strukturierten Ablage von Informationen, über im Alltag genutzte Dateiformate bis zu komplexen, automatisierten Publikationsszenarien.
- ➔ XML-Technologien lassen sich bereits mit einfachen Mitteln studieren und ausprobieren.
- ➔ Um letztlich attraktive Medienprodukte zu produzieren, ist neben den theoretischen Grundlagen viel Entwicklerpraxis nötig (z. B. durch Bearbeitung entsprechender Bachelor- und Masterthemen).