Anleitung.docx.zip – Streifzug durch die Welt der Containerformate

tekom-Jahrestagung 2016 – Stuttgart, 09. November

Dr. Thomas Meinike

Hochschule Merseburg | FB Wirtschaftswissenschaften und Informationswissenschaften
web.hs-merseburg.de/~meiniket/
thomas.meinike@hs-merseburg.de

Motivation

- Wir gehen im beruflichen und privaten Alltag mit einer Vielzahl an Dokumenten- und Medienformaten um.

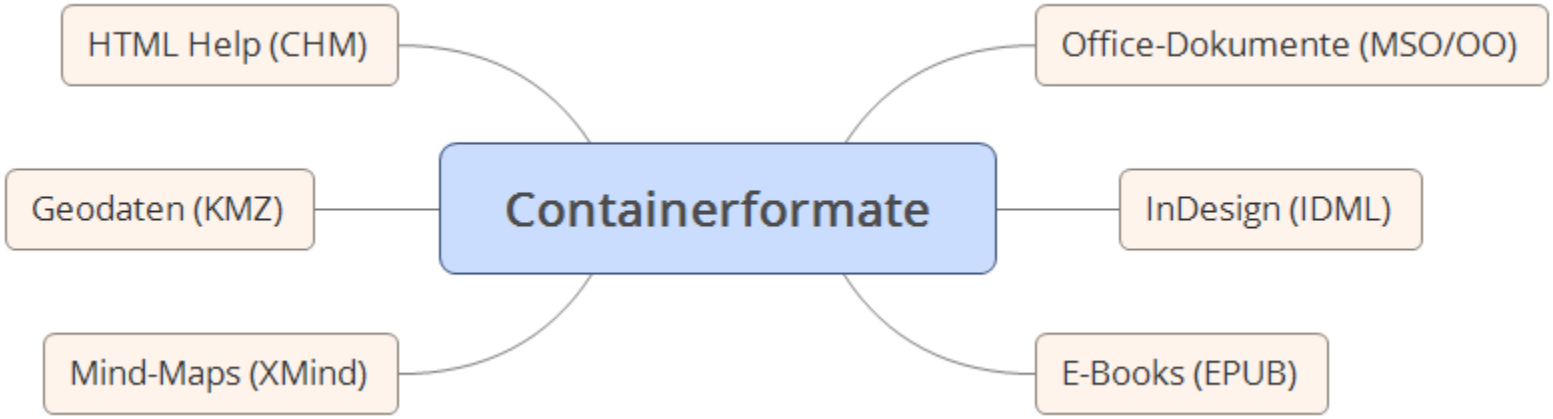
- Zum Erstellen oder Konsumieren mit geeigneter Software sind Detailkenntnisse des jeweiligen Datei-Innenlebens kaum von Bedeutung.

- In den letzten Jahren haben sich Containerformate etabliert, häufig ZIP-komprimierte Archive mit einer weiter verzweigten Verzeichnis- und Dateistruktur (z. B. die EPUB-Version des Tagungsbandes).

- Der Vortrag vermittelt Einblicke zum Aufbau typischer Formate aus dem Umfeld der Technischen Kommunikation und stellt praktische Ansätze zu ihrer Produktion auf der Basis von XML-Technologien vor.

Überblick

➔ Im Kern werden diese Einsatzbereiche behandelt:



➔ Theoretische Details lassen sich nur anreißen, die Demo-Anwendungen und der mitgelieferte Sourcecode dienen zur Inspiration und Vertiefung.

ZIP-Format 1/2

- Ende der 1980er Jahre von Phil Katz entwickelt (DOS-Nutzer werden sich an die Programme pkzip.exe / pkunzip.exe erinnern).

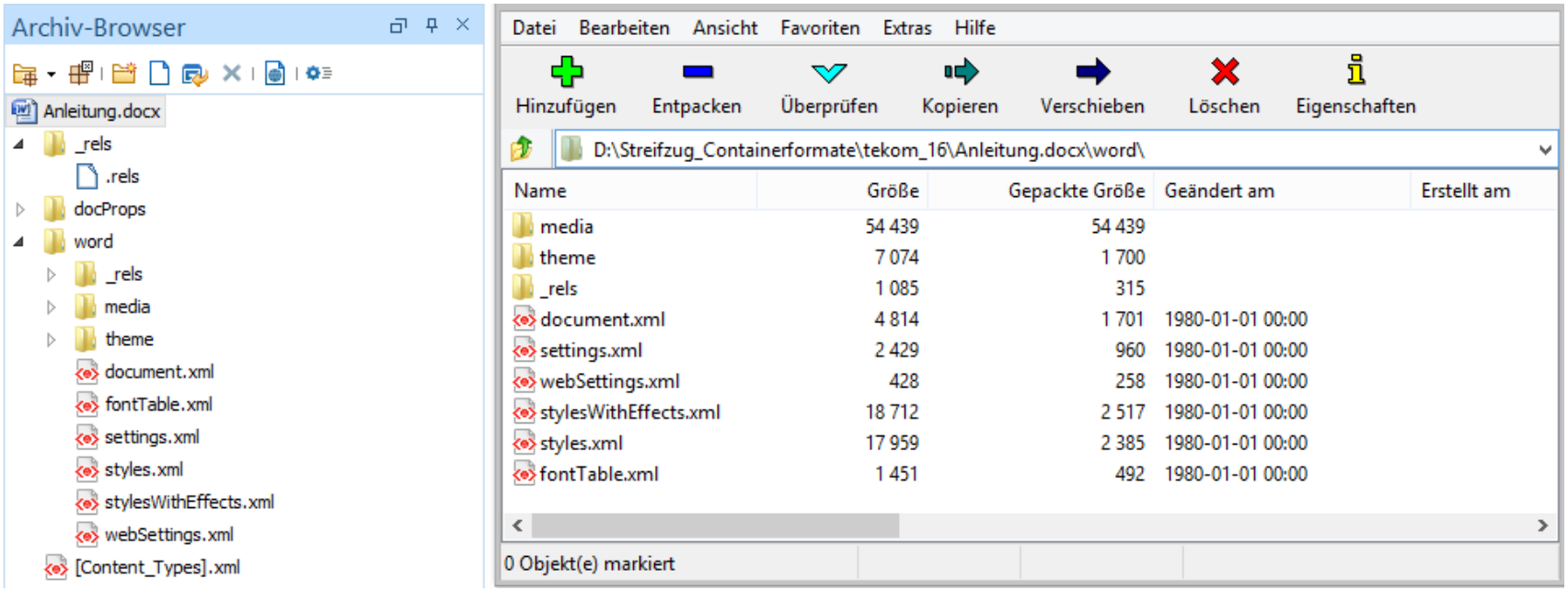
- Die beiden ersten Bytes eines ZIP-Archivs sind auf **PK** gesetzt (die Initialen des Entwicklers), auch eine DOCX-Datei von Microsoft Word weist diese Kennung auf:

	0001	0203	0405	0607	0809	0A0B	0C0D	0E0F	0123456789ABCDEF
00000	504B	0304	1400	0600	0800	0000	2100	05EB	PK.....!..ë
00010	7825	9B01	0000	4506	0000	1300	0802	5B43	x%›...E.....[C
00020	6F6E	7465	6E74	5F54	7970	6573	5D2E	786D	ontent_Types].xm
00030	6C20	A204	0228	A000	0200	0000	0000	0000	l ¢..(.....
00040	0000	0000	0000	0000	0000	0000	0000	0000	.....

- (Aktuelle) Word-Dokumente sind also ZIP-Archive!

ZIP-Format 2/2

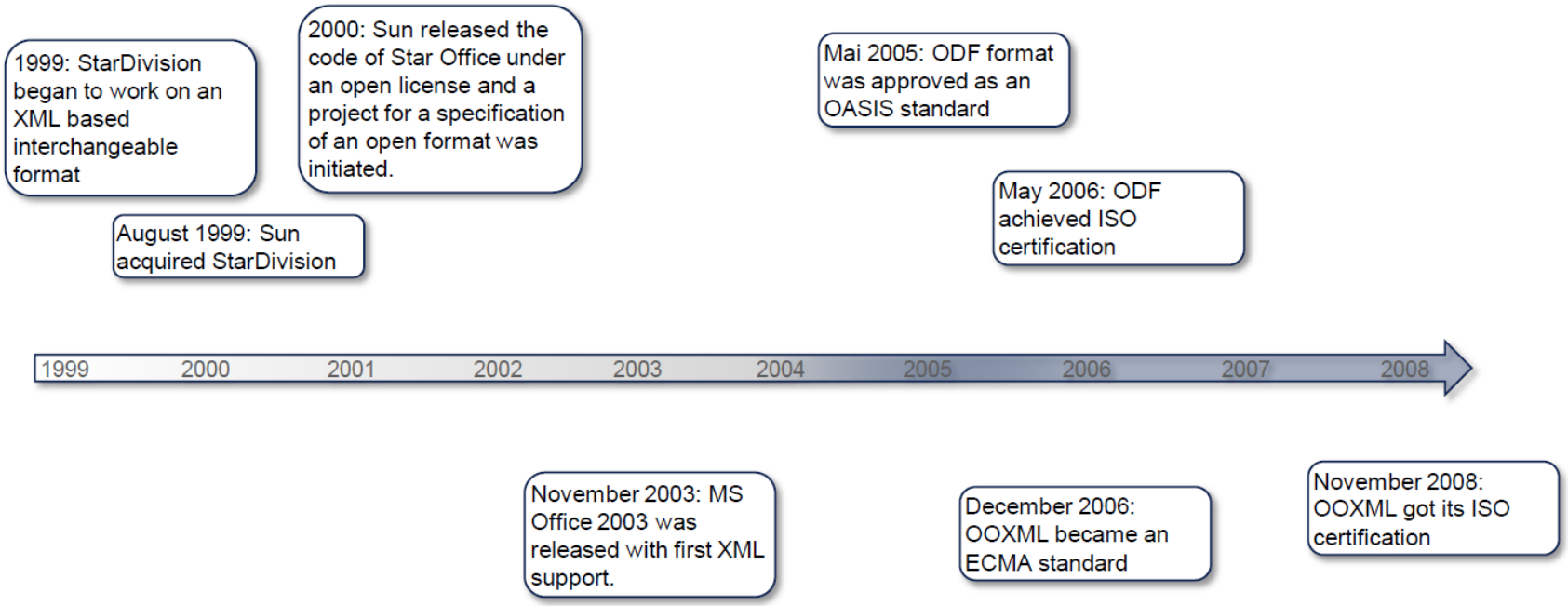
➔ Eine DOCX-Datei lässt sich in *.docx.zip umbenannt oder direkt mit ZIP-Werkzeugen öffnen und die interne Struktur wird ersichtlich:



links: Archiv-Browser im <oXygen/> XML Editor | rechts: File Manager von 7-Zip

Office-Formate

➔ Aktuelle Formate von Microsoft Office und OpenOffice / LibreOffice sind einerseits ZIP-Archive und andererseits inhaltlich XML-basiert.



Bildquelle zur zeitlichen Entwicklung: Simon Trang, Uni Göttingen 2010
dbis.informatik.uni-goettingen.de/Teaching/Seminars/ML-WS0910/XMLbasedDocFormats-simon-trang.pdf

Office Open XML (OOXML) 1/7

- Von Microsoft mit Office 2007 eingeführt und als offizielle Standards ECMA-376 (2006) bzw. ISO/IEC 29500 (2008) verabschiedet.

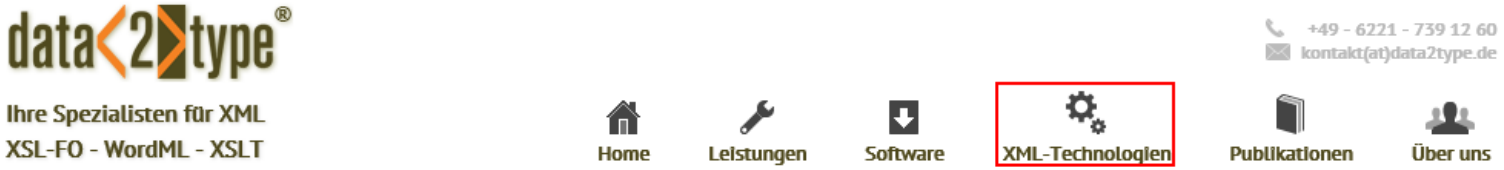
- Die Standard-Dokumente beschreiben die XML-Syntax für jegliche Inhalte sowie Formatierungen von Excel, PowerPoint und Word.

Bildquelle: Jirka Kosek
xmlguru.cz/2007/07/czech-comments-ooxml

Office Open XML (OOXML) 2/7

- Der Standard definiert die Paketstrukturen der ZIP-Archive sowie spezielle Auszeichnungssprachen für die Vokabulare:
  - WordprocessingML / WordML (Word)
  - PresentationML (PowerPoint)
  - SpreadsheetML (Excel)

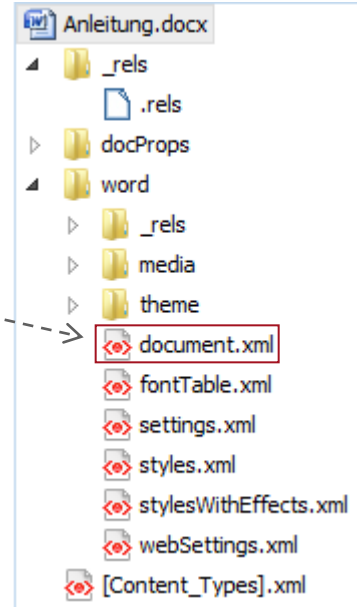
- Hilfreiche Zusammenfassungen zu diesen und weiteren Sprachen liefert data2type.de → XML-Technologien:



Office Open XML (OOXML) 3/7 → Word

➔ Wesentliche Dateien im DOCX-Archiv (zur Synthese anpassen):

- docProps/app.xml (Eigenschaften)
- docProps/core.xml (Metadaten)
- word/document.xml (Inhalte)
- word/media (Bilderverzeichnis)
- word/_rels/document.xml.rels
  (interne Referenzen: Bilder, Links, Styles)
- fontTable.xml, numbering.xml, styles.xml, …
  (aus Dateinamen ersichtliche Funktionen)



Office Open XML (OOXML) 4/7 → Word

➔ Grobstruktur von document.xml:

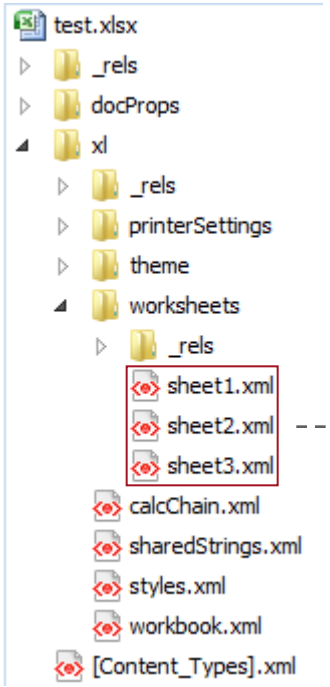
```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document xmlns="[Office-Namensräume …]">
  <w:body>
    <w:p w:rsidR="00BA60FC" w:rsidRDefault="00A20935" w:rsidP="00A20935">
      <w:pPr>
        <w:pStyle w:val="berschrift1"/>
      </w:pPr>
      <w:r>
        <w:t>Überschrift</w:t>
      </w:r>
    </w:p>
    <w:p w:rsidR="00A20935" w:rsidRPr="00A20935" w:rsidRDefault="00A20935"
         w:rsidP="00A20935">
      <w:r w:rsidRPr="00A20935">
        <w:t xml:space="preserve">Ein </w:t>
      </w:r>
      <w:r w:rsidRPr="00A20935">
        <w:rPr>
          <w:rPr>
            <w:b/>
          </w:rPr>
          <w:t>Absatztext</w:t>
        […]
    </w:p>
  </w:body>
</w:document>
```

Einige WordML-Elemente:

w:b = fett
w:p = Absatzbereich
w:pPr = Absatzformate
w:pStyle = Formatzuweisung
w:rPr = Eigenschaften für Fließtext
w:r = Fließtextbereich (run)
w:t = Textbereich im Fließtext

Office Open XML (OOXML) 5/7 → Excel

➔ SpreadsheetML – XLSX-Struktur:



Die eigentlichen Daten liegen unterhalb von xl/worksheets in sheet1…n.xml.

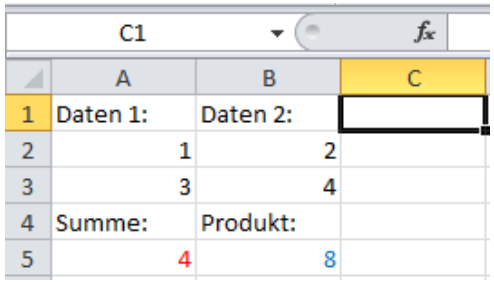
➔ Weitere Details lassen sich mit konkreten Dokumenten erkunden.

Office Open XML (OOXML) 6/7 → Excel

➔ SpreadsheetML – Werte und Formeln:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<worksheet ...>
  <dimension ref="A1:B5"/>[…]
  <sheetData>
    <row r="1" spans="1:2" x14ac:dyDescent="0.25">
      <c r="A1" t="s"><v>1</v></c>
      <c r="B1" t="s"><v>3</v></c>
    </row>
    <row r="2" spans="1:2" x14ac:dyDescent="0.25">
      <c r="A2"><v>1</v></c>
      <c r="B2"><v>2</v></c>
    </row>
    <row r="3" spans="1:2" x14ac:dyDescent="0.25">
      <c r="A3"><v>3</v></c>
      <c r="B3"><v>4</v></c>
    </row>
    <row r="4" spans="1:2" x14ac:dyDescent="0.25">
      <c r="A4" t="s"><v>0</v></c>
      <c r="B4" t="s"><v>2</v></c>
    </row>
    <row r="5" spans="1:2" x14ac:dyDescent="0.25">
      <c r="A5" s="1"><f>SUM(A2,A3)</f><v>4</v></c>
      <c r="B5" s="2"><f>B2*B3</f><v>8</v></c>
    </row>
  </sheetData>[…]
</worksheet>
```

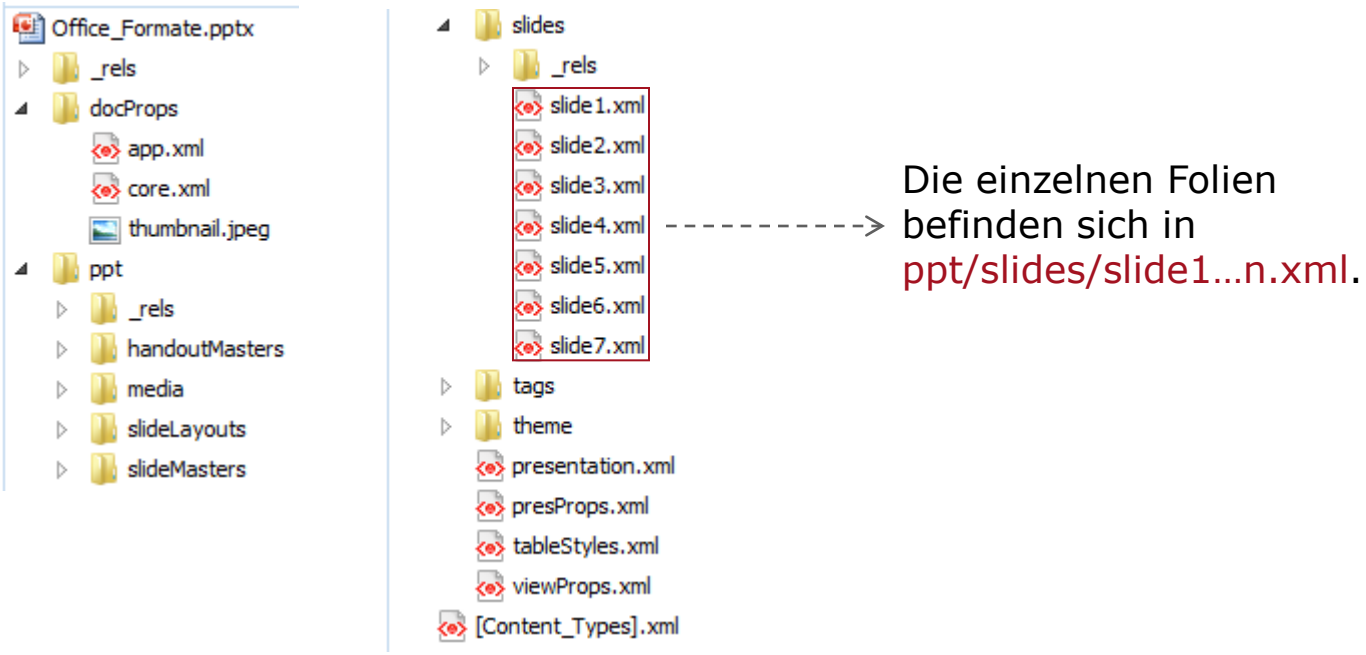
xl/worksheets/sheet1.xml



Zellen-Texte enthält xl/sharedStrings.xml.

Office Open XML (OOXML) 7/7 → PowerPoint

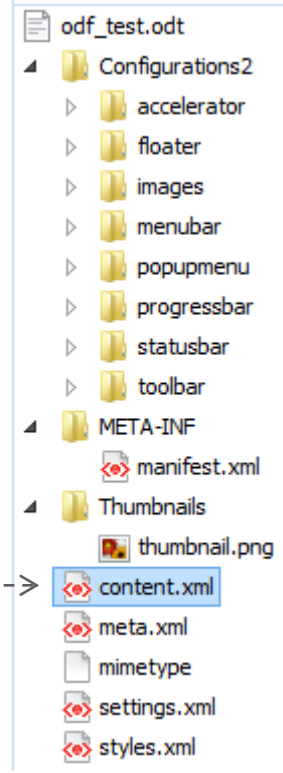
➔ PresentationML – PPTX-Struktur:



➔ Weitere Details lassen sich mit konkreten Dokumenten erkunden.

OpenDocument kompakt

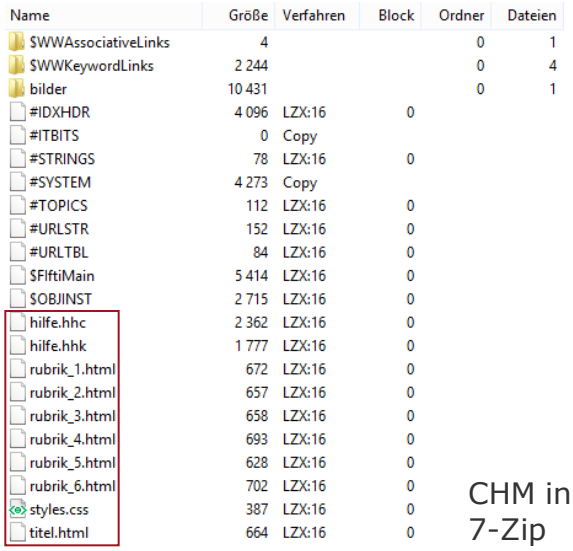
- Entwickelt von Sun, dann von OASIS übernommen, 2006 als ISO/IEC 26300 standardisiert.
- Formate von OpenOffice (Writer, Calc, Impress) und weiteren kompatiblen Anwendungen wie LibreOffice.
- Ebenfalls in ZIP-Archiven organisiert.
- Hauptinhalte von ODT (Writer) liegen in content.xml.
- Die Strukturen von ODS und ODP sind ähnlich aufgebaut.



- Weitere Details lassen sich mit konkreten Dokumenten erkunden.

HTML Help (CHM)

- Löste das in den 1990er Jahren unter Windows verwendete HLP-Format ab (eingeführt mit Windows 98 / Internet Explorer 4.0).
- Spezifische Struktur, die mit dem HTML Help Workshop (HWS) angelegt wird (Projekt: name.hhp, Inhaltsverzeichnis: name.hhc, Index: name.hhk). Hinzu kommen HTML-Topics, CSS und Bilder.
- Kompression mit dem LZX-Algorithmus.
- Die relevanten Dateien sind Text-basiert, also mittels XSLT generierbar.
- Sofern der HWS installiert ist, reicht zur Erzeugung der Hilfecompiler hhc.exe aus.

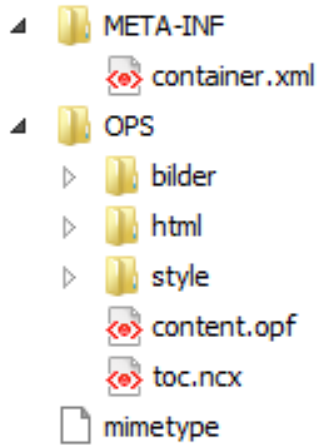


Name	Größe	Verfahren	Block	Ordner	Dateien
$WWAssociativeLinks	4			0	1
$WWKeywordLinks	2 244			0	4
bilder	10 431			0	1
#IDXHDR	4 096	LZX:16	0		
#ITBITS	0	Copy			
#STRINGS	78	LZX:16	0		
#SYSTEM	4 273	Copy			
#TOPICS	112	LZX:16	0		
#URLSTR	152	LZX:16	0		
#URLTBL	84	LZX:16	0		
$FlftiMain	5 414	LZX:16	0		
$OBJINST	2 715	LZX:16	0		
hilfe.hhc	2 362	LZX:16	0		
hilfe.hhk	1 777	LZX:16	0		
rubrik_1.html	672	LZX:16	0		
rubrik_2.html	657	LZX:16	0		
rubrik_3.html	658	LZX:16	0		
rubrik_4.html	693	LZX:16	0		
rubrik_5.html	628	LZX:16	0		
rubrik_6.html	702	LZX:16	0		
styles.css	387	LZX:16	0		
titel.html	664	LZX:16	0		

CHM in 7-Zip

E-Books (EPUB)

- Offenes Format, Herausgeber IDPF (2007: EPUB 2, 2011: EPUB 3).
- EPUB 2.0.1 noch sehr verbreitet, vor allem auf E-Ink-Lesegeräten.
- Im ZIP-Archiv sind XHTML-Inhalte, CSS, Bilder und Fonts sowie Struktur- und Steuerdateien (mimetype, container.xml, content.opf, toc.ncx) enthalten (→ Editor Sigil).
- Sämtliche Text-basierten Inhalte lassen sich mittels XSLT erzeugen (→ epubMinFlow).
- EPUB 3 setzt auf (X)HTML5-Erweiterungen, u. a. vereinfachte Navigation (toc.ncx → nav.html).

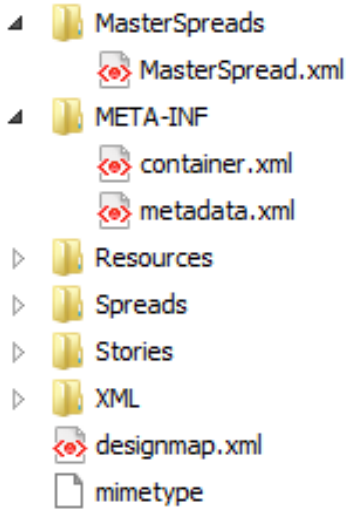


InDesign Markup Language (IDML)

- Von Adobe ab InDesign CS4 (2008) als Export- bzw. Austauschformat für die Kompatibilität zwischen ID-Versionen konzipiert.

- Enthält im ZIP-Container alle Inhaltsdaten und die zugehörigen Formatierungen, zur Synthese sind hauptsächlich zu erstellen:
  - META-INF/metadata.xml (Metadaten)
  - designmap.xml (Referenzen / IDs)
  - Spreads/Spread_*.xml (Seitenaufteilung)
  - Stories/Story_*.xml (Seiteninhalte)
  - [+ Grundgerüst mit Vorgaben]

- Wesentliche Dateien mittels XSLT produzierbar, interessant zur Umsetzung von Massendaten in feste Layouts.

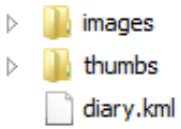


Sonstiges (Auswahl)

- **Mind-Maps** – Programme wie FreeMind verwenden zur Datenablage XML-Formate, XMind speichert zudem in ZIP-komprimierter Form:

Siehe Mind-Map im Überblick (Folie 3).

- **Geoformate** – Für die Ablage von via GPS gewonnenen Koordinaten wird ebenfalls XML eingesetzt. Google Earth etablierte das KML-Format, gepackt als KMZ.

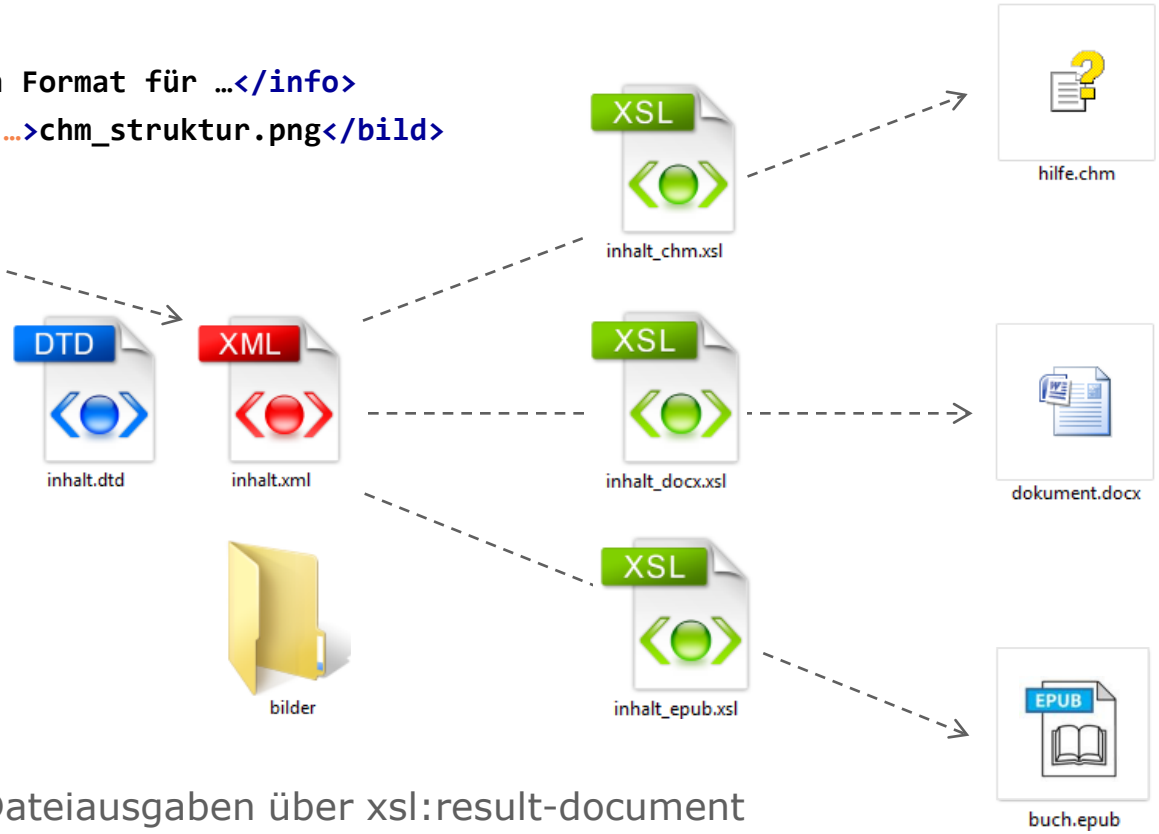


- **SVGZ** – Ermöglicht die weitere Reduktion der Vektordaten von SVG-Dokumenten. Verwendet wird die GZip-Kompression u. a. von Adobe Illustrator und Inkscape (*.svgz).

Praktische Demonstration 1/5

➔ **Ausgabe von CHM, DOCX und EPUB** aus einer XML-Datenstruktur nach dem Single-Source-Prinzip mit XSLT 2.0 [1]:

```
[…]<rubrik name="HTML Help">
  <info>HTML Help (CHM) ist ein Format für …</info>
  <bild abbtext="CHM-Struktur" …>chm_struktur.png</bild>
</rubrik>[…]
```



[1] Dateiausgaben über xsl:result-document

Praktische Demonstration 2/5

➔ CHM-Produktion

- Batchjob process_chm.cmd:

✓ Ausgabeverzeichnisse anlegen

✓ Bilder vorkopieren

✓ XSLT: HTML-Topics und CSS

✓ XSLT: HTML Help-spezifisch
name.hhp, name.hhc, name.hhk

✓ Compileraufruf: hhc name.hhp

Praktische Demonstration 3/5

➔ DOCX-Produktion

- Batchjob process_docx.cmd:

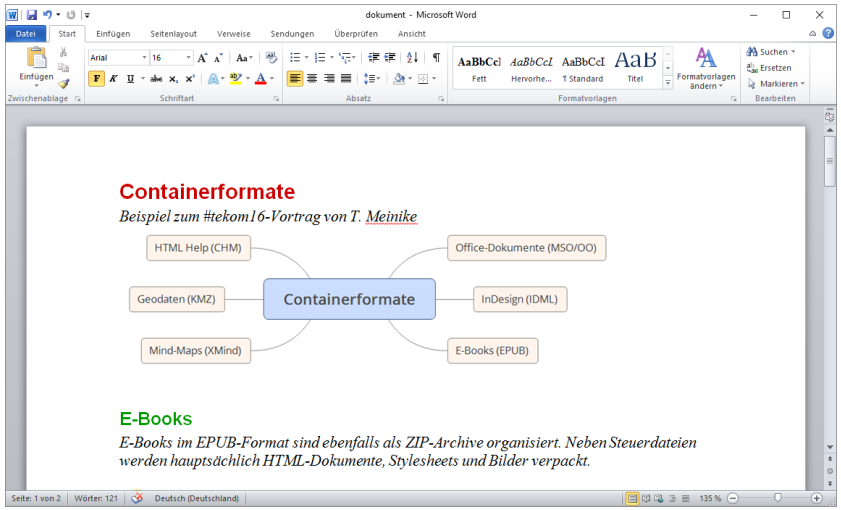
✓ Ausgabeverzeichnisse anlegen

✓ Basisstruktur vorkopieren

✓ Bilder vorkopieren (word/media)

✓ XSLT: docProps/core.xml, word/_rels/document.xml.rels, word/document.xml

✓ Ausgabestruktur packen



Praktische Demonstration 4/5

➔ EPUB-Produktion

- Batchjob process_epub.cmd:

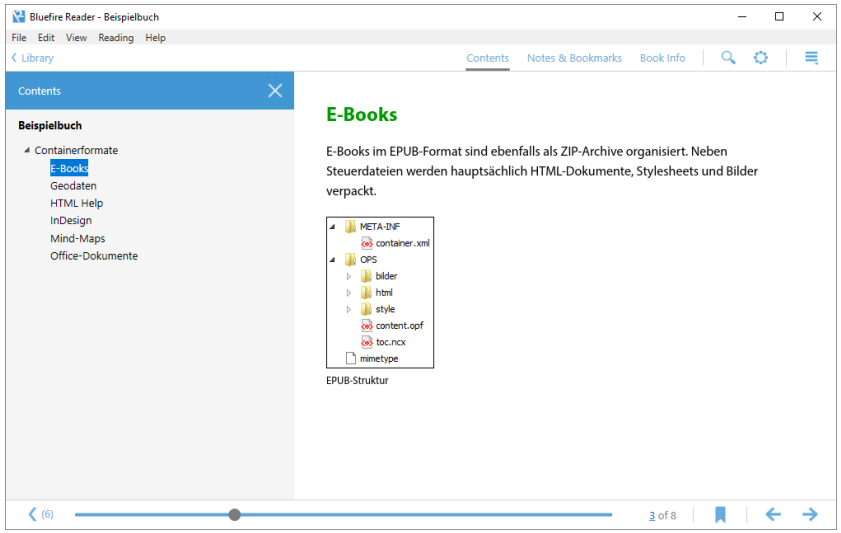
✓ Verzeichnisse anlegen

✓ Bilder vorkopieren

✓ XSLT: HTML-Kapitel, CSS

✓ XSLT: EPUB 2-spezifisch
mimetype, container.xml
content.opf, toc.ncx

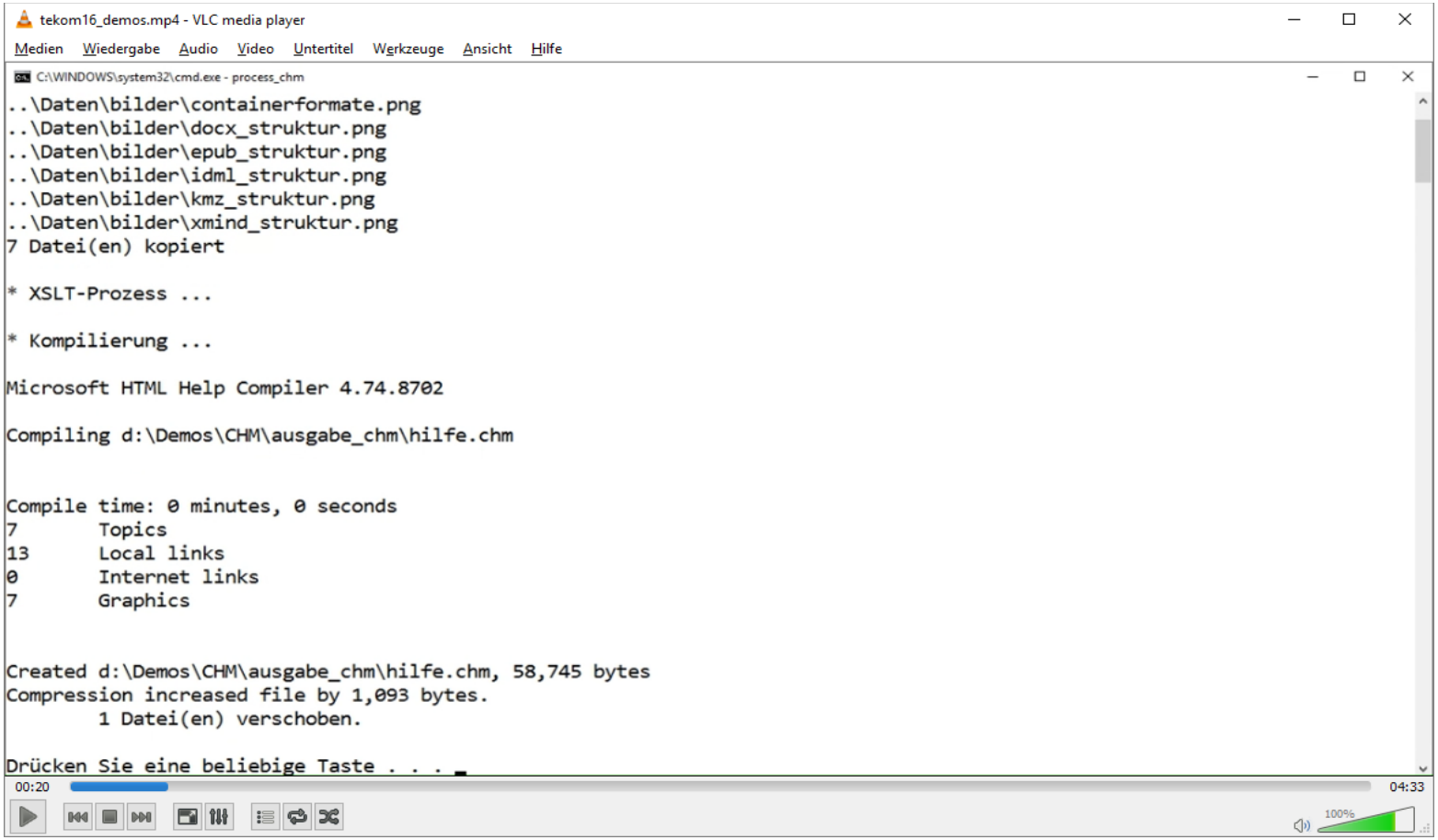
✓ Ausgabestruktur packen
und mit EpubCheck prüfen



HOME HOCHSCHULE MERSEBURG FH
University of Applied Sciences

Praktische Demonstration 5/5

➔ **Ablaufvideo** (mit dem Code als Download verfügbar)



Fazit und Ausblick

- Containerformate werden vielfältig eingesetzt.

- Die Kenntnis ihres Innenlebens ermöglicht die Produktion unabhängig von den eigentlichen Anwendungsprogrammen im Rahmen von XML-Workflows.

- Für den Alltag ohne produktive Ambitionen mit XSLT & Co. kann sich das Gelernte ebenfalls lohnen, z. B. zur Rettung von Inhalten defekter Office-Dokumente:
  name.docx → name.zip → word/media → Bilder extrahieren.

Literatur und Ressourcen

Adobe: IDML File Format Specification; adobe.com/content/dam/Adobe/en/devnet/indesign/cs55-docs/IDML/idml-specification.pdf
DateiEndung.com: Dateiendung .svgz; dateiendung.com/format/svgz
ECMA International: TC45 – Übersicht über Office Open XML; ecma-international.org/news/TC45_current_work/OpenXML_White_Paper_German.pdf
Foltin, C. et al.: FreeMind; freemind.sourceforge.net
Hendricks, K. und Massay, D.: EPUB-Editor Sigil; sigil-ebook.com
International Digital Publishing Forum (IDPF); idpf.org
Meinike, T.: epubMinFlow (2010); datenverdrahten.de/epubMinFlow
Meinike, T.: Einfach publizieren und benutzen – EPUB-Format in Theorie und Praxis; Entwickler Magazin 4.10, S. 99-106
Meinike, T.: XSLT-Programmierung – effektiv und schmerzfrei!; Tagungsband zur Jahrestagung 2011, S. 313-315
Microsoft: HTML Help Downloads; msdn.microsoft.com/de-de/library/windows/desktop/ms669985(v=vs.85).aspx
OASIS: Open Document Format for Office Applications (OpenDocument) Version 1.2; docs.oasis-open.org/office/v1.2/OpenDocument-v1.2.pdf
Pavlov, I.: 7-Zip; 7-zip.org
SyncRO Soft SRL: <oXygen/> XML Editor; oxygenxml.com
The Apache Software Foundation: OpenOffice; openoffice.org
The Document Foundation: LibreOffice; libreoffice.org
Wikipedia: Keyhole Markup Language; en.wikipedia.org/wiki/Keyhole_Markup_Language
Wikipedia: LZX (algorithm); en.wikipedia.org/wiki/LZX_(algorithm)
Wikipedia: Zip (file format); en.wikipedia.org/wiki/Zip_(file_format)
XMind Ltd.: XMind; xmind.net

Feedback erwünscht …

➔ URL direkt aufrufen (auch nach der Tagung möglich):
in08.honestly.de

➔ Oder QR-Code scannen:

Danke für Ihr Interesse!