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

Dr. Thomas Meinike, Hochschule Merseburg

Motivation

In der Technischen Kommunikation besonders ausgeprägt, aber auch im Alltag von „normalen“ Anwendern wird mit einer Vielzahl von Datei- und Dokumentenformaten umgegangen. Dabei kommt es beim Erstellen mit einschlägiger Software oder schlichtem Medienkonsum nicht auf Detailkenntnisse ihres Innenlebens an. Dennoch ist es zumindest erhellend, einen tieferen Einblick zu wagen und je nach Anforderung die gewonnenen Erkenntnisse auch produktiv zu nutzen. Im Mittelpunkt stehen typische Containerformate, womit überwiegend ZIP-gepackte Archive mit einer weiter verzweigten Verzeichnis- und Dateistruktur gemeint sind.

Einstieg

Das ZIP-Format hat sich als Quasi-Standard zum verlustfreien Packen von Dateien etabliert. Es wurde Ende der 1980er-Jahre von Phil Katz entwickelt [1]. Schaut man sich eine Datei mit der Erweiterung .zip in einem Hex-Editor an, sind am Anfang die Initialen PK des Entwicklers zu erkennen (in den ersten beiden Bytes als 504B kodiert). Auch beim Betrachten einer Word-Datei mit der Erweiterung .docx zeigt sich gemäß Abbildung 1 dieses Indiz für ein ZIP-Archiv.

```
        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  ................
```

Abb. 1: Anfang einer DOCX-Datei in Hex-Ansicht

Eine solche Datei kann problemlos in .docx.zip oder nur .zip umbenannt und mit geeigneten Werkzeugen geöffnet werden. Abbildung 2 zeigt links eine Word-Datei im Archivbrowser des XML-Editors <oXygen/> [2] und daneben im File-Manager von 7-Zip [3]. Die interne Struktur ist bereits ansatzweise zu erkennen, insbesondere eine Reihe von XML-Dokumenten. Das Umbenennen ist bei diesen Werkzeugen nicht nötig.

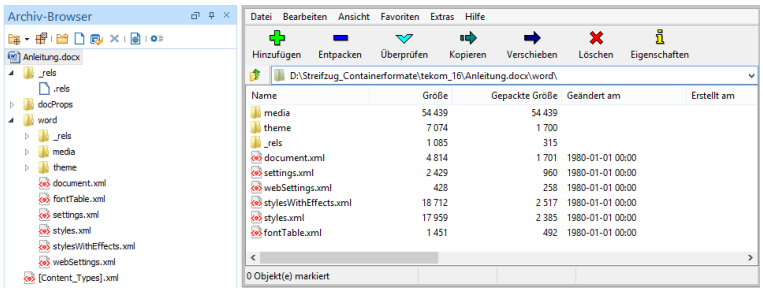

Abb. 2: DOCX-Datei im <oXygen/> XML Editor und in 7-Zip

Ausgewählte Formate im Überblick

Office-Dokumente

Das genannte Word-Format (.docx) ist aus der Office-Suite von Microsoft bekannt. Auch die von Excel (.xlsx) und PowerPoint (.pptx) unmittelbar gespeicherten Formate sind ZIP-Archive und enthalten spezifische XML-Inhalte und ggf. weitere Medien wie Bilddateien. Während Word den eigentlichen Textinhalt in document.xml ablegt, sind die Excel-Tabellen hauptsächlich in sheet{1…n}.xml zu finden, während PowerPoint-Folien als slide{1…n}.xml hinterlegt sind. Zusätzliche Dokumente beschreiben u. a. Metadaten, Formatierungen und Verweise.

Diese Formate wurden von Microsoft mit der Office-Version 2007 eingeführt und als Office Open XML (OOXML) standardisiert (ECMA-376 sowie ISO/IEC 29500). Einen kompakten Einblick in die mehrere tausend Seiten umfassenden Spezifikationen bietet [4].

Microsoft hat OOXML entworfen, obwohl es die Chance gegeben hätte, ebenfalls auf den bereits seit 2005 vorhandenen OASIS-Standard OpenDocument [5] zu setzen. Darauf basierende Formate für Texte, Tabellen und Präsentationen werden von OpenOffice [6] bzw. Derivaten wie LibreOffice [7] verwendet (Erweiterungen .odt, .ods, .odp). In den ZIP-Archiven sind wiederum Verzeichnisse sowie XML- und Mediendateien zu finden.

E-Books

Das populäre EPUB-Format ist als eine Art Website im ZIP-Gewand konzipiert. Seine Standardisierung obliegt dem IDPF-Gremium [8]. Aktuell sind die Versionen 2 (2007) und 3 (2011) gebräuchlich. Inhaltlich werden XHTML-Dokumente, Stylesheets, Bilder, Audio- und Video-Dateien (EPUB 3) sowie Navigations- und Metadaten verpackt. Der Artikel unter [9] vermittelt die wesentlichen Grundlagen ausgehend von EPUB 2. Abbildung 3 zeigt den Editor Sigil [10], der direkt auf der Buchstruktur arbeitet.

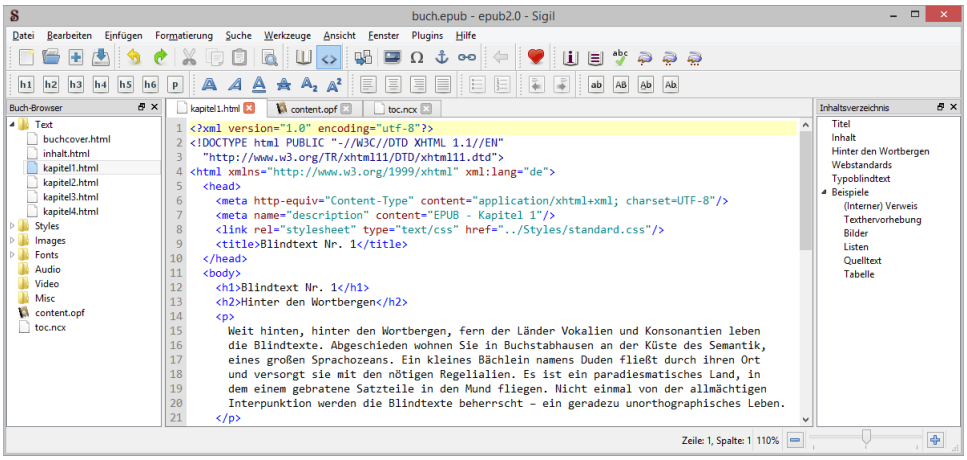

Abb. 3: EPUB-Editor Sigil – Archivstruktur links

Weitere für E-Books auf Kindle-Geräten konzipierte Formate wie MOBI und KF8 verfolgen ähnliche Ansätze.

InDesign-Austauschformat

Die Firma Adobe führte mit InDesign CS4 (2008) das Austauschformat IDML (InDesign Markup Language) für die Interoperabilität mit anderen InDesign-Versionen bzw. zum Erstellen von Layout-Dokumenten mit externen Werkzeugen ein. Es handelt sich ebenfalls um eine ZIP-Architektur, die in Analogie zu den Office-Formaten layoutorientierte XML-Dokumente (u. a. Spreads und Stories) organisiert [11].

Online-Hilfen

Das noch weit verbreitete, aus den 1990er Jahren stammende und von Microsoft im Umfeld von Windows 98 und dem Internet Explorer 4.0 eingeführte CHM-Format liegt ebenfalls als komprimiertes Archiv vor. Die Kompression erfolgt mit dem LZX-Algorithmus [12]. Enthalten sind je eine Projektdatei, ein Inhaltsverzeichnis, ein Stichwortverzeichnis und die eigentlichen Inhalte (HTML, CSS, Bilder). Das genannte Werkzeug 7-Zip kann auch mit CHM-Dateien umgehen.

Sonstiges

Im Kontext der Technischen Kommunikation sind auch Mind-Maps und zugehörige Anwendungen von Interesse. Gespeichert werden häufig XML-Formate entweder direkt (z. B. von FreeMind [13]) oder wiederum in ZIP-Archiven verpackt (z. B. von XMind [14]), siehe Abbildung 4.

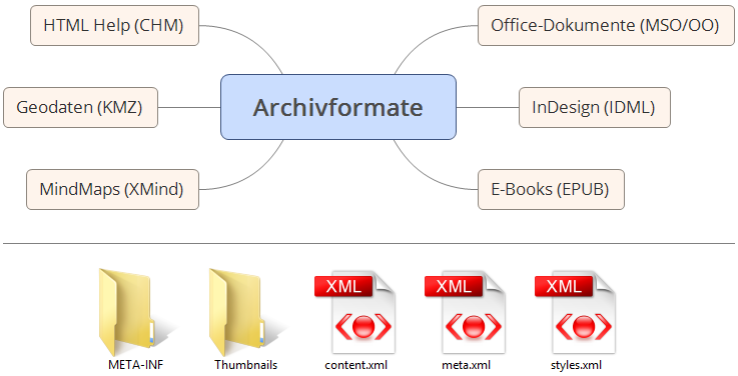

Abb. 4: XMind-Map und interner Aufbau

Bei der Gewinnung von Geodaten mit GPS-nutzenden Geräten fallen ebenfalls spezielle XML-Formate wie das für Google Earth entwickelte KML an, gepackt als KMZ [15]. Erwähnenswert ist noch das platzsparende SVGZ-Format [16] für Vektorgrafiken. Zeichenwerkzeuge bieten üblicherweise eine separate Option zur GZip-komprimierten Ablage.

Produktive Nutzung und Ausblick

Mit Detailkenntnissen der genannten Formate lassen sich diese auch erzeugen. Die XML-Basis bildet den gemeinsamen Nenner für Lösungen nach dem Single-Source-Publishing-Prinzip. Dazu sind insbesondere Transformationen mittels XSLT in die Zielstrukturen und -inhalte geeignet. Ansätze zur Erstellung von CHM, DOCX, EPUB und IDML hat der Autor bereits beschrieben und mit Code untersetzt [17, 18]. Im Vortrag werden die Interna der Formate und ihre Produktion weiter vertieft.

Literaturangaben und Links

[1] Wikipedia: Zip (file format); https://en.wikipedia.org/wiki/Zip_(file_format)

[2] SyncRO Soft SRL: <oXygen/> XML Editor; http://www.oxygenxml.com/

[3] Pavlov, I.: 7-Zip; http://www.7-zip.org/

[4] ECMA International: TC45 – Übersicht über Office Open XML; http://www.ecma-international.org/news/TC45_current_work/OpenXML_White_Paper_German.pdf

[5] OASIS: Open Document Format for Office Applications (OpenDocument) Version 1.2; http://docs.oasis-open.org/office/v1.2/OpenDocument-v1.2.pdf

[6] The Apache Software Foundation: OpenOffice; http://www.openoffice.org/

[7] The Document Foundation: LibreOffice; http://www.libreoffice.org/

[8] International Digital Publishing Forum (IDPF): http://idpf.org/

[9] Meinike, T.: Einfach publizieren und benutzen – EPUB-Format in Theorie und Praxis, Entwickler Magazin 4.2010, S. 99–106; http://web.hs-merseburg.de/~meiniket/PDF/EM/EM_4.10_Meinike_EPUB.pdf

[10] Hendricks, K. und Massay, D.: Sigil; https://sigil-ebook.com/

[11] Adobe: IDML File Format Specification; https://www.adobe.com/content/dam/Adobe/en/devnet/indesign/cs55-docs/IDML/idml-specification.pdf

[12] Wikipedia: LZX (algorithm); https://en.wikipedia.org/wiki/LZX_(algorithm)

[13] Foltin, C. et al.: FreeMind; http://freemind.sourceforge.net/

[14] XMind Ltd.: XMind; http://www.xmind.net/

[15] Wikipedia: Keyhole Markup Language; https://en.wikipedia.org/wiki/Keyhole_Markup_Language

[16] DateiEndung.com: Dateiendung .svgz; http://www.dateiendung.com/format/svgz

[17] Meinike, T.: epubMinFlow – Ein minimaler Workflow zur automatisierten Umsetzung von E-Books im EPUB-Format (2010); http://datenverdrahten.de/epubMinFlow/

[18] Meinike, T.: XSLT-Programmierung – effektiv und schmerzfrei! In: tekom, Gesellschaft für technische Kommunikation e. V., Tagungsband zur Jahrestagung 2011, S. 313–315 / Material unter: http://web.hs-merseburg.de/~meiniket/vortraege.php

für Rückfragen:
thomas.meinike@hs-merseburg.de