E-Books – bitte mit Format!

EPUB-Produktion, Content Management-System, XML-Struktur, mediengerechtes Layout, Lesesysteme, Metadaten

www.dpc-consulting.de

Fabian Kern ist freier Berater und Trainer für digitales Publizieren. Mit **digital publishing competence** ist er spezialisiert auf E-Books, Mobile Apps, Enhanced E-Books und Web-Anwendungen. Er berät Verlage und Medienunternehmen bei der Produktion von modernen Digitalmedien und bietet Seminare und Fortbildungen in diesem Bereich an. Neben seinem Schwerpunkt auf Projektconsulting ist er Dozent bei der Akademie des Deutschen Buchhandels und an der LMU München.

Seitdem Amazon mit ihrem Kindle-Ökosystem einen Marktboom für das digitale Lesen von E-Books auf eReadern, Tablets und Smartphones ausgelöst hat, müssen sich auch Verlage, die bisher auf medienneutrale Datenhaltung in der Produktion verzichtet haben, mit der effizienten Erzeugung von Digitalmedien auseinandersetzen. Als Dateistandard für E-Books hat sich in Markt und Technik mittlerweile das EPUB-Format durchgesetzt. Das vom Amazon eingesetzte Konkurrenzformat KF8 kann aus EPUB über eine automatisierte Konvertierung abgeleitet werden.

Zur Erzeugung von EPUB-E-Books gibt es mittlerweile eine breite Palette von Werkzeugen – von einfachen Konverter-Plug-ins für Microsoft Word über den Export aus DTP-Anwendungen wie Adobe InDesign bis zu dedizierten Enterprise-Produktionsumgebungen wie etwa PXE von Aptara. Ein hoch effizienter und nahezu beliebig skalierbarer Weg zur E-Book-Produktion ist daneben aber die Erzeugung aus einem XML-basierten Content Management-System.

Aufbau und Inhalte eines EPUB-E-Book

EPUB ist ein verhältnismäßig einfach aufgebautes Dateiformat, das auf offenen Webstandards basiert – für eine automatisierte Konvertierung aus XML-Daten ein unschätzbarer Vorteil. Eine EPUB-Datei ist technisch gesehen nichts anderes als ein ZIP-Container mit einem durch den Standard festgelegten internen Dateisystem. Neben einer Abfolge von HTML-Dokumenten für das Vorhalten der Text-Inhalte eines E-Book-Produktes enthält eine EPUB-Datei typischerweise folgende Datenstrukturen:

- Produktmetadaten wie Autor, Titel und ISBN sowie eine Manifest-Sektion mit einer Auflistung aller Dateien im ZIP-Container und ihrer Mimetype-Angaben. Die Angaben werden in

einer EPUB-spezifischen XML-Struktur, der OPF-DTD, vorgehalten.
- Für den Aufbau eines hierarchischen Inhaltsverzeichnisses durch eReader-Anwendungen wird die Gliederungsstruktur in einer EPUB-spezifischen XML-Struktur eingebettet, der NCX-DTD.
- Formatierung und Layout der Inhalte über ein eingebettetes CSS-Stylesheet: Gängige E-Book-Reader unterstützen dafür ein vereinfachtes Subset der Eigenschaften von CSS 2.1.
- Alle weiteren notwendigen Ressourcen wie Bilddaten, eingebettete Schriftarten: Dazu gibt der EPUB-Standard eine übersichtliche Liste zugelassener Medientypen vor.

Eine EPUB-Datei ist insofern letztlich eine Art sehr einfach aufgebaute, statische Website, die für die Distribution an Kunden und Lese-Anwendungen in eine ZIP-Datei verpackt wird. Diese Struktur eignet sich optimal für Konvertierungsprozesse auf Basis medienneutraler XML-Daten – unabhängig davon, ob das eingesetzte CMS einen Export bereits selber zu Verfügung stellt oder ob dafür eine eigene Implementierung notwendig ist. Welche Funktionen und Prozesse sind aber erforderlich, um alle Komponenten eines EPUB-E-Books erzeugen zu können? Welche Anforderungen müssen an ein CMS gestellt werden, um eine effiziente und reibungslose Verarbeitung sicherzustellen?

Die Dokument-Zusammenstellung zur E-Book-Publikation

Im Optimalfall werden die zugrunde liegenden XML-Dokumente im CMS möglichst atomar und in feingranularer Form adressierbar vorgehalten, d.h. ein bereits vorliegendes Print-Produkt sollte z.B. nicht als monolithische Quell-Datei vorliegen. Als sinnvolle Container-Struktur für die XML-Daten bieten sich ▶

beispielsweise die Hauptkapitel einer Print-Publikation an, bei Produkttypen wie Lexika oder Periodika auch die einzelnen Artikel. Daneben ist es für die Erzeugung der notwendigen HTML-Dokumente ideal, wenn jede inhaltlich sinnvolle Struktur in den Daten über IDs adressierbar ist und separat in eine Publikation integrierbar ist.

Für den Aufbau einer EPUB-Datei sollten das CMS oder die Konvertierungs-Schicht einen Mechanismus vorsehen, mit dem die zu integrierenden Inhalte als Datei und/oder XML-Container referenziert und so zu einer Publikation zusammengestellt werden können. Dabei sollte es nicht nur möglich sein, die Referenzen sequentiell anzuordnen. Auch die Integration von Hierarchien und Zwischenebenen, eine Verschachtelung von Teilen unterschiedlicher Quelldokumente und eine Adressierungsmöglichkeit für beliebige Inhalte-Container geben einem System die notwendige Flexibilität, um auch komplexe Publikationen und Neuzusammenstellung bestehender Inhalte zu E-Books zu realisieren.

Vom XML-Dokument zum E-Book-Content

Als HTML-Dialekt in EPUB wird XHTML verwendet, allerdings werden innerhalb von E-Books gegenüber dem modernen Webdesign üblicherweise stark vereinfachte HTML-Strukturen verwendet. Die Transformation XML/HTML ist dabei eine klassische Down-Konvertierung. Beim Aufsetzen einer solchen Konvertierung hat es sich bewährt, die Quell-DTD zunächst nach der typographischen Rolle und Semantik der XML-Elemente zu analysieren und auf dieser Basis ein Mapping von XML-Elementen auf HTML-Elemente vorzunehmen. Optimalerweise werden aufgrund der XML-Semantiken der Quell-DTD bereits zugehörige CSS-Klassen erzeugt, deren Namensraum auf der einen Seite auf die Formatierungsanforderungen des E-Book abgestimmt ist, auf der anderen Seite möglichst einfache Rückschlüsse auf die XML-Quellen erlaubt, falls dies aus Qualitätssicherungs-Gründen notwendig ist.

Typischerweise wird beim Design einer solchen Transformation mit einer Konvertierungstabelle gearbeitet, die abhängig vom Element-Typ die XML-Elemente auf HMTL-Elemente abbildet. Elemente mit Überschriften-Charakter werden dabei etwa zu HTML-Header-Elementen, Absatz-artige Elemente zu Paragraph-Elementen, Inline-Elemente zu Span-Elementen. Für Container-Elemente wie Textkästen, Kapitelstrukturen und ähnliches wird als Blockelement gewählt. Auf der Basis einer so aufgebauten Arbeitsanweisung für die Element-Transformation wird nicht nur ein Programmierer in der Lage sein, eine effiziente Konvertierung zu entwickeln, etwa mit XML-eigenen Mechanismen wie der Transformations-Sprache XSLT, auch die Gestaltung der E-Books für ein ansprechendes Layout wird deutlich vereinfacht.

Gestaltungsvorgaben und CSS-Design für E-Books

Für eine mediengerechte Gestaltung von E-Book-Content spielen vor allem zwei Faktoren eine zentrale Rolle:

01 Das Prinzip des „Reflow-Layout“: Als zentrales Feature für den Leser bieten nahezu alle Lesesysteme die Möglichkeit, Gestaltungsmerkmale wie Schriftarten und Schriftgrößen zur Laufzeit zu verändern. Ein eReader gestaltet dazu den Seitenumbruch flexibel neu, ein E-Book-Layout darf insofern keine Merkmale verwenden, die sich auf fixe Gestaltungsmerkmale, feste Positionierungen und eine vorhersehbare Seitengröße verlassen.

   Ein E-Book-CSS sollte insofern für Eigenschaften wie Schriftgrößen, Abstände und Seitenaufbau ausschließlich relative Angaben wie em- oder Prozent-Werte verwenden. Bei der Umsetzung eines existierenden Print-Layout auf eine

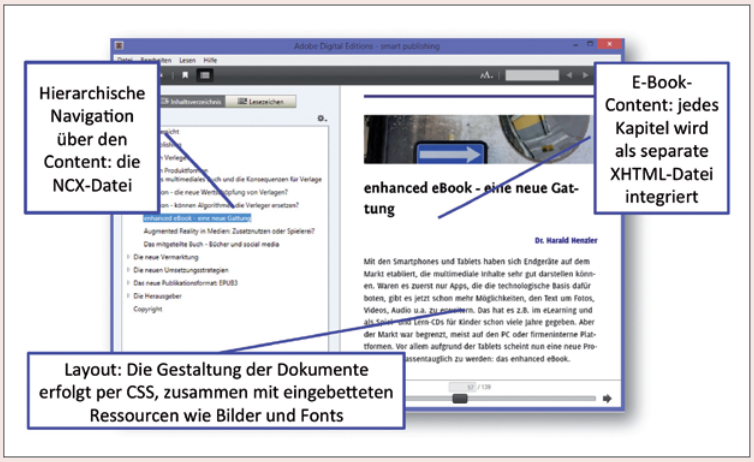

Bild 1: E-Books aus der Sicht des Kunden

E-Book-Gestaltung sind dabei insofern oft auch umfangreiche konzeptionelle Überlegungen notwendig, um sowohl Wiedererkennbarkeit einer Gestaltung wie auch breite Kompatibilität sicherzustellen.

02 Die Beschränkungen und die Heterogenität der Lesesysteme: Gängige eReader-Anwendungen unterstützen in der Regel nur einen Bruchteil der Gestaltungsmöglichkeiten, die im modernen Webdesign gang und gäbe sind. Dazu unterscheiden sich die Lesesysteme – ähnlich wie Webbrowser in früheren Jahren – zum Teil erheblich in ihrer Interpretation der CSS-Angaben und in ihrer Unterstützung von Webstandards.

In der Praxis sind E-Book-Stylesheets deswegen meist ein Kompromiss zwischen ansprechendem Layout, Kosten/Nutzen-Erwägungen für gestalterische Besonderheiten und möglichst breiter Kompatibilität zwischen den verschiedensten Lesegeräten und E-Book-Applikationen.

Gerade jedoch für die Produktion von großen Mengen an Publikationen auf Basis medienneutraler Daten ist diese Art des Designs durchaus von Nutzen: Einmal aufgesetzt, kann ein Layout nahezu beliebig wiederverwendet werden. Bei Bedarf kann ein Basis-Layout um Customizing-Schichten wie CSS-Varianten für bestimmte Produktreihen oder Genres von Publikationen ergänzt werden.

Was noch zur Publikation gehört: Bilder, Ressourcen, Metadaten

Sinnvollerweise verwaltet das CMS, auf dessen Basis produziert wird, auch die notwendigen Media Assets wie Bilder oder etwa eingebettete Fonts. Bei Export bzw. Konvertierung sollte das System selbstverständlich prüfen, ob alle Ressourcen vorhanden und vollständig sind. Ein Mapping von zugelassenen Dateitypen auf die notwendigen Mimetype-Angaben ist für dabei unerlässlich. Gleichzeitig sollten Bilddaten bereits in den gängigen Dateitypen für Web-Grafiken vorliegen oder systemseitig konvertiert sowie in Web-optimierte Größen und Auflösungen skaliert werden können.

Für den Aufbau der Produktmetadaten in der EPUB-Datei sind minimal Angaben zu Autor, Titel, Verlag und ISBN erforderlich. Je nach Menge, Publikationsrythmus und Komplexität des Produktstamms können diese Angaben entweder in einer Konfigurationsschicht zur Konvertierung vorgehalten werden oder, falls notwendig, per Schnittstelle aus Verlagsanwendungen für die Materialstamm-Verwaltung abgefragt werden.

Produktionsprozess, Qualitätssicherung, Implementierungswege

Für den Produktionsprozess auf Basis dieser Voraussetzungen sind mindestens folgende Schritte erforderlich, die in einem Export oder einer Konvertierung implementiert werden müssen:

- Export aller XML-Dokumente abhängig von den Strukturvorgaben für den Produktaufbau, Konvertierung von XML-DTD in die vorgegebenen HTML-Strukturen
- Export aller notwendigen Bilder und eingebetteten Ressourcen, ggf. unter Konvertierung in geeignete Dateitypen, Größen und Auflösungen
- Auswertung der Strukturvorgaben für den Produktaufbau und der erzeugten HTML-Strukturen für den Aufbau der EPUB-Navigation in der NCX-Datei
- Auswertung der erzeugten Dateien und Zusammenstellung von Dateipfaden, IDs und Mimetype-Angaben zur Manifest-Sektion in der OPF-Datei
- Export der Produktmetadaten als XML-Struktur und Integration in die OPF-Datei

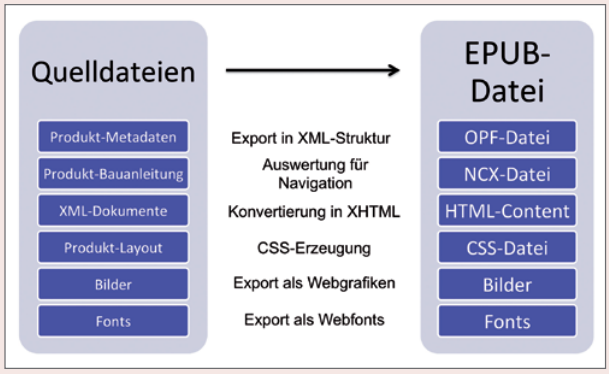

Bild 2: E-Books aus der Sicht des CMS

- Verteilung der erzeugten Dateien in ein Template-basiert erzeugtes Dateisystem für den EPUB-ZIP-Container
- Verpacken und Benennen der EPUB-Datei nach den Benennungskonventionen des Verlages, ggf. unter Beachtung von Vorgaben der E-Book-Distributoren

Für die Behandlung von Fehlern und Sonderfällen gibt es neben den systemseitig bereits zur Verfügung stehenden Parsern und QS-Routinen der Content Management-Systeme noch zwei zentrale Werkzeuge, die unbedingt verwendet werden sollten:

- Validierung der HTML- und CSS-Dokumente über die Web-Anwendungen, die vom W3C zur Verfügung gestellt werden.
- Validierung der erzeugten EPUB-Datei über das vom IDPF zur Verfügung gestellte Epubcheck-Tool. Dabei werden alle internen Dateistrukturen und Referenzen innerhalb der EPUB-Datei auf Konsistenz geprüft, lästige Reklamationen von Distributoren, E-Book-Shops und Kunden können so weitgehend vermieden werden.

Für die Implementierung einer EPUB-Konvertierung stehen vielfältige Werkzeuge zur Verfügung: Für die Transformation von XML nach HTML bietet sich zunächst der Einsatz von XSLT an. Neben dedizierten XSLT-Engines können aufgrund der zumeist relativ einfachen Konvertierungslogik auch die XML-Parser und XSLT-Bibliotheken vieler Programmiersprachen in Verbindung mit Konfigurationstabellen für die Steuerung dieser zentralen Komponente verwendet werden.

Zur Steuerung des Konvertierungsablaufs und für das notwendige Datei-Handling wird es daneben sinnvoll sein, auch eine prozedurale Komponente zu entwickeln, die die Schritte von Datenbank-Export, Transformation, Erzeugung des Dateisystems und des EPUB-Containers übernimmt und etwaige Fehler und Ausnahmen behandelt. Je nach eingesetztem CMS und verwendeter Produktionsumgebung bieten sich dazu nahezu alle modernen Hochsprachen an, die ausreichend mit Bibliotheken für Datenbank-Zugriff, XML-Handling und das Ansprechen von Dritt-Anwendungen ausgerüstet sind, etwa Java, PHP oder C#.

Vorteile einer CMS-basierten E-Book-Produktion

Gegenüber einem manuellen oder halbautomatisierten Prozess für die EPUB-Erzeugung, wie er zum Beispiel aus DTP-Anwendungen wie InDesign heraus realisiert wird, bietet eine CMS-basierte E-Book-Produktion die klassischen Vorteile von XML-Anwendungen: Einheitlichkeit der Ausgabe auf Basis medienneutraler Datenstrukturen, Wiederverwendbarkeit der Quelldaten für unterschiedlichste Produktkombinationen, hohe Skalierbarkeit und effiziente Produktion auch bei großen Datenmengen und besonderen Anforderungen an Durchsatz und Zeitverhalten.

Diese klassischen Vorzüge einer CMS-Anwendung kommen für die E-Book-Produktion insbesondere dann zur Geltung, wenn bereits ein derartiges System im Einsatz ist und große Mengen an Bestandsdaten existieren. Das Aufsetzen eines EPUB-Exports bietet dann für Content-Anbieter die Möglichkeit, mit einer überschaubaren einmaligen Investition in die Produktionsumgebung nahezu beliebige Volumen an E-Books zu realisieren.