Outputformate: Alle sind gut, aber keines ist besser

Lieber PDF oder AFP? Wann XML, wann XSL-FO? Und was ist überhaupt mit HTML5? Dieser Artikel sowie das Glossar auf der nächsten Doppelseite klären über Vor- und Nachteile der gängigen Formate im Outputmanagement auf.

Kommt die Sprache auf Datenformate im Outputmanagement, wird es schnell kontrovers. Dabei ist die Diskussion, welches Format nun besser ist, müßig – denn alle haben ihre Existenzberechtigung; seitenbezogene Formate wie AFP und PDF genauso wie inhaltsorientierte à la HTML und XML. Entscheidend ist vielmehr die Frage, welche Prämissen ein Unternehmen hat und wie die Strukturen seiner Dokumentenverarbeitung aussehen. Für einen Betrieb, der täglich nur ein paar Hundert Seiten druckt, gibt es keinen Grund, sich beispielsweise mit AFP zu beschäftigen. Anders die Situation bei Unternehmen mit einem Druckvolumen von mehreren Millionen Seiten täglich. Es wird wohl um AFP nicht herumkommen, denn bekanntlich gilt das einst von IBM entwickelte Format als der Standard für den hochvolumigen und sicheren Produktionsdruck. Zu Recht, denn es besitzt Features, die andere Formate nicht haben (siehe auch Glossar auf Seite 34). So ist beispielsweise die Überwachung des Drucks bei AFP am ausgereiftesten: Werden Inhalte nicht korrekt oder unvollständig ausgegeben, erfolgt automatisch eine Fehlermeldung. Nicht umsonst findet sich AFP vor allem in der industriellen Produktion von Rechnungen, Kontoauszügen, Wertpapier-/Depotaufstellungen, Überweisungsbelegen, Versicherungspolicen etc.

AFP für den sicheren Massendruck in hoher Qualität

Außerdem besitzt AFP ein ausgeprägtes Management für die Darstellung auf mehreren Seiten, für die Schachtsteuerung sowie für den Simplex-/Duplexdruck – also für Bereiche, die in der Massenproduktion unerlässlich ist. Wegen der umfassenden und flexiblen Ressourcenverwaltung und Kompaktheit des Datenstroms ist AFP in Anwenderkreisen äußerst beliebt. Produkt- und Anwendungsentwickler schätzen dessen durchdachte und gut dokumentierte Architektur. Selbst ernst zu nehmende Alternativen wie PDF erreichen nicht diese Qualität beim Druck – auch wenn man es mit der Spezifikation PDF/VT versucht hat. Doch es bleibt nur ein Kompromiss, die Vorzüge von AFP für den Massendruck und die äußerst hohe Flexibilität von PDF zu einem neuen Format zu bündeln.

Andererseits ist PDF mit seinen verschiedenen Spezifikationen aufgrund seiner hohen Kompatibilität ein weltweit anerkannter Standard und hat sich zweifelsohne als das Format für die langfristige, revisionssichere Archivierung (PDF/A) und die Erstellung von barrierefreien Dokumenten (PDF/UA) durchgesetzt. Letztlich hängt die Entscheidung, ob man in AFP oder PDF drucken soll, von der konkreten Situation ab. Wer als Unternehmen beispielsweise Dokumente in hoher Stückzahl im Original-Layout archivieren will oder muss, könnte sich überlegen, diese in PDF bzw. PDF/A auszugeben – er würde sich damit eine notwendige Konvertierung von AFP nach PDF ersparen.

Viele Datenformate gibt es, die Dokumentenfluten verarbeiten können.

PDF und HTML5 sind keine Konkurrenten

Sowohl AFP als auch PDF orientieren sich rigoros an A4 als Seitenstandard – und kommen deshalb dann zum Einsatz, wenn man seine Dokumentenverarbeitung stark

AFP ist das gängigste Format für den Massendruck von Transaktionspost.

© Nomad_Soul – Fotolia

an der A4-Ausgabe ausrichtet. Für die Darstellung im Web und auf mobilen Endgeräten sind beide Formate aber ungeeignet. Hier kommt HTML5 ins Spiel. Der W3C-Standard ist derzeit das intelligenteste Format für die größen- und kanalunabhängige Erstellung und Ausgabe von Dokumenten.

Es ermöglicht die Re-Formatierung, beispielsweise von A4 zum Smartphone-Display, die Konvertierung von seitenbezogenen in textorientierte Formate, die Extraktion von Einzeldaten (u.a. Rückgewinnung von Rechnungspositionen) und den Aufbau von Inhaltsverzeichnissen und Indexlisten. Mehr noch: Mit HTML5 lassen sich auch audiovisuelle Elemente, Weblinks und Charts einbetten. So entstehen auf diese Weise nicht nur multikanalfähige, sondern auch intelligente Dokumente, die dem Nutzer einen über die reine Textdarstellung hinausgehenden Mehrwert bieten.

Die Entwicklung von HTML5 kommt funktional einem Quantensprung gleich. Die neue Version gilt mittlerweile als die „Sprache des Webs" und kann mit relativ geringem Aufwand ohne weiteres auch als Druckformat benutzt werden. Leider existiert immer noch die irrige Annahme, dass HTML5 und PDF „Konkurrenten" seien, vor allem, wenn es um die Hinterlegung von Strukturinformationen geht. Weit gefehlt, denn schließlich ist HTML5 der kleinste gemeinsame Nenner für die kanalunabhängige Darstellung und Ausgabe von Dokumenten. PDF wird also nicht verschwinden, im Gegenteil: Beide Formate bedingen einander. So kann HTML5 innerhalb der Dokumentenverarbeitung die Vorstufe zu PDF sein, denn für bestimmte Prozesse wie Archivierung wird nach wie vor PDF/A benötigt.

Schluss mit den „Religionskriegen"!

Geht es um die Etablierung eines kanalübergreifenden Outputmanagements, landet man früher oder später bei einem weiteren Format, das zunehmend an Bedeutung gewinnt: XSL-FO (siehe Glossar). Die auf XML basierende Auszeichnungssprache besitzt gegenüber HTML einen entscheidenden Vorteil: Sie ermöglicht nicht nur eine von der Seitengröße unabhängige Erstellung und Ausgabe von Dokumenten, sondern liefert auch eine Vielzahl ausgefeilter Funktionen für die Gestaltung von Seiten. Mit XSL-FO ist es möglich, hochwertige Druckerzeugnisse zu erzeugen. Anders als HTML, das sich vor allem für Browser-Anwendungen eignet, kommt XSL-FO eher im Druck- und Archivierungsbereich zum Einsatz, also dort, wo innerhalb eines Dokuments viele Seiten anfallen.

Bliebe noch XML: Die nach ISO normierte Auszeichnungssprache gilt mittlerweile als Standard für die Übergabe von Daten aus Fachanwendungen an die Outputinstanz eines Unternehmens. XML-Technologien sind heute derart ausgereift, dass es für die Datenextraktion keiner besonderen Softwarekomponenten bedarf. Das gilt auch für die anderen Formate. Ob nun AFP, PDF, HTML5 oder XSL-FO – die gängigen Standards eines modernen Outputmanagements besitzen inzwischen einen so hohen Abdeckungsgrad durch IT-Lösungen, dass es für ein Unternehmen kein Problem sein sollte, eine Gesamtarchitektur, die alle Szenarien bedient, zu entwickeln und zu etablieren; zumal die Kosten dafür auch überschaubar sind.

Daher sollte man endlich aufhören, „Religionskriege" um das beste Format zu führen. Es geht um eine grundsätzliche Entscheidung, nämlich darum, wie die Dokumentenverarbeitung eines Unternehmens strategisch ausgerichtet wird: Welche Kommunikationskanäle werden künftig in welcher Intensität eine Rolle spielen? Mit welchem Dokumentenaufkommen ist zu rechnen? Wie wird sich das Verhältnis zwischen physikalischem und elektronischem Versand entwickeln? Erst wenn diese Fragen beantwortet sind, weiß man auch, welche Formate an welcher Stelle zum Tragen kommen. Alle besitzen sie Stärken, aber auch Schwächen. Entscheidend sind die Anwendungsszenarien, denn sie allein bestimmen die Relevanz der Formate für das jeweilige Outputmanagement eines Unternehmens.

Harald Grumser, Compart

Weitere Informationen:
www.compart.com

In der elektronischen Kommunikation läuft HTML5 dem PDF vielleicht bald den Rang ab.

© mornius – Fotolia

Glossar

Die wichtigsten Datenformate in der Dokumentenverarbeitung – ein Überblick

AFP

Advanced Function Presentation (AFP) ist eine Software- und Hardwarearchitektur für den Datenstrom zum Erstellen, Formatieren, Betrachten, Suchen, Drucken und Verteilen von Informationen für eine breite Auswahl an Druckern und Ausgabegeräten. Für die Bildschirmanzeige ist ein spezieller Viewer erforderlich. Unterschieden werden grundsätzlich zwei Datenströme:

- AFPDS (AFP Datenstrom) beziehungsweise MO:DCA (Mixed Object: Document Content Architecture). AFPDS ist ein Datenstrom, der Dokumente druckerunabhängig und seitenweise beschreibt. Typischerweise enthalten AFPDS-Dateien Zehntausende von Dokumenten (Rechnungen, Kontoauszüge, Briefe etc.).
- IPDS (Intelligent Printer Data Stream) ist der Datenstrom, der von Druckserver-Programmen verwendet wird, um IPDS-Drucker zu steuern. In der Regel druckt man AFPDS-Datenströme auf IPDS-Druckern.

Verbreitung:

AFP ist vor allem für große Mengen an Dokumenten und für hohe Druckgeschwindigkeiten bis 3000 Seiten pro Minute gedacht. In der industriellen Produktion von Rechnungen und Kontoauszügen ist AFP das am meisten verbreitete Format. AFP zeichnet sich durch eine exzellente Trennung von Nutz- und wiederkehrenden Daten (u.a. bei Formularen) aus. Ursprünglich für den digitalen Schwarzweißdruck konzipiert und dort verbreitet, ist AFP inzwischen auch im Farbdruck als Standard etabliert.

Eine AFP-Druckdatei mit 5000 Seiten beispielsweise enthält neben dem Text für 5000 Empfänger auch Verweise auf Ressourcen (Fonts, Grafiken usw.), die nur einmal abgelegt werden. Archivierungssysteme beherrschen oft den originalen AFP-Druckstrom einschließlich Ressourcenverwaltung oder konvertieren beispielsweise eine AFP-Druckdatei in 5000 einzelne PDF-Dateien.

PDF (Portable Document Format)

Das Portable Document Format (PDF) ist ein plattformunabhängiges Dokumentenformat und wurde in den 1990er Jahren mit folgenden Zielen entwickelt:

- Austausch und Darstellung elektronischer Dokumente
- Texte und Bilder unabhängig von der Auflösung grafisch darstellen
- Dokumente für die (Web-)Ansicht optimieren
- interaktive Funktionen anbieten

PDF ist eng verwandt mit PostScript, im Unterschied zu diesem aber keine Programmiersprache, sondern ein Dateiformat. Es beschreibt die Seiten eines Dokuments mittels Objekten und der dazugehörigen Strukturinformationen. Hinzu kommen interaktive Elemente wie Formulare, Bookmarks, Soundobjekte und Thumbnails, die auf diese Weise nur in elektronischen Dokumenten wiedergegeben werden können. Um eine Darstellung auch auf Ausgabegeräten mit kleiner Anzeigefläche (Smartphones etc.) zu optimieren, können in einem PDF so genannte Auszeichnungen (ähnlich HTML-Tags) eingelagert werden. Diese ermöglichen ein Umbrechen der Seiteninhalte, was allerdings von fast keinem PDF-Viewer unterstützt wird. Zudem erleichtern sie das Konvertieren des Inhalts in andere Formate sowie das Vorlesen des Dokuments für sehbehinderte Nutzer mit einem speziellen Programm (Barrierefreiheit).

PDF-Dateien beschreiben das erzeugte Layout in einer vom Drucker und von Voreinstellungen unabhängigen Form weitgehend originalgetreu – einer der wesentlichen Unterschiede zu Beschreibungs- und Auszeichnungssprachen wie SGML und HTML bezüglich einer hohen Layout-Treue.

Verbreitung und Vorteile:

PDF-Dokumente sind weltweit in nahezu allen Marktsegmenten zu finden und wegen der zahlreichen Vorteile sehr beliebt. Dazu zählen:

- Kompatibilität: PDF ist plattformunabhängig, das heißt, ein PDF-Dokument, das beispielsweise mit einer Windows-Applikation erstellt wurde, lässt sich anschließend auf einem Unix-Server verarbeiten und auf einem Mac-Rechner anschauen.
- Extrafunktionen: Das PDF-Format baut auf der Seitenbeschreibungssprache PostScript auf, bietet darüber hinaus den direkten Zugriff auf Seiten sowie Features für Verschlüsselung, Komprimierung, interaktive Navigation etc.
- Industriestandard: PDF hat sich zum international anerkannten ISO-Standard für den elektronischen Austausch von Dokumenten entwickelt. Zudem ist PDF heute das am meisten eingesetzte Format für die Produktion von Druckvorlagen in der digitalen Druckvorstufe.
- Viewer/Reader: Einer der Hauptgründe für die hohe Akzeptanz des Formats ist, dass es für alle gängigen Plattformen kostenlose Anzeige-Programme Viewer/ Reader) gibt. Das bedeutet, dass sich auf jeder Hardware- und Softwarearchitektur der Inhalt einer PDF-Datei ohne grafischen Unterschied darstellen lässt.

PDF-Standards:

Aufgrund der hohen Akzeptanz von PDF als Dokumentenformat haben sich daraus verschiedene Spezifikationen für bestimmte Bereiche der Dokumentenerzeugung und -verarbeitung entwickelt. Dazu gehören unter anderem:

- PDF/A für die kompakte, langfristige, elektronische Archivierung – ohne viel Aufwand und gesetzeskonform (Compliance). Seit 2005 hat sich das Format als ISO-Standard etabliert.
- PDF/VT für die Erstellung/Anzeige und den Druck von Transaktionsdokumenten (Kontoauszüge, Rechnungen, Lieferscheine).
- PDF/UA zur Erstellung von PDF-Dateien, die auch für Nutzer mit Sehbehinderungen oder motorischen Störungen zugänglich sind. Ziel dieses noch zu etablierenden ISO-Standards ist es, festzulegen, wie PDF-Dokumente und die darin enthaltenen Informationselemente (u.a. Grafiken, Texte, Multimedia, Formularfelder) bereit gestellt werden müssen, damit diese auch von Menschen mit Behinderung gelesen und bearbeitet werden können (Barrierefreiheit).

HTML5

HTML5 ist eine textbasierte Auszeichnungssprache zur Strukturierung und semantischen Beschreibung von Dokumenten. Sie findet bereits breite Anwendung, vor allem auf mobilen Geräten und im Internet. Das Besondere des neuen Formats: Es bietet zahlreiche Funktionalitäten für Grafik (2D-/3D-Grafiken) und Multimedia (Audio/Video), die von bisherigen Standards wie HTML 4.01 und XHTML nicht direkt unterstützt werden. Nützlich an HTML5 ist auch die Einbettung von Webfonts. Damit lassen sich mit dem Browser auch „Hausschriften“ von einem Server herunterladen. Unterstützt ein Browser die HTML5-Schriftarten nicht, werden sie durch einen Standardfont wie Arial oder Verdana ersetzt.

XML (Extensible Markup Language)

XML ist eine vom World Wide Web Consortium (W3C) nach ISO normierte Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Text-

daten. XML wird unter anderem für den plattformunabhängigen Austausch von Daten zwischen Computersystemen eingesetzt, insbesondere über das Internet. Das W3C definiert XML als eine Metasprache, auf deren Basis durch strukturelle und inhaltliche Parameter anwendungsspezifische Sprachen definiert werden. Dazu gehören unter anderem XSL-FO (vor allem im Druck-/Archivierungsbereich), XHTML (webbasierte Applikationen) und SVG.

In XML werden alle Elemente streng strukturiert, vergleichbar mit der aus der Windows-Welt bekannten Menübaumstruktur. Dabei bestimmt stets ein einleitendes Schlüsselwort (Start-Tag) und ein abschließendes Keyword (End-Tag) die jeweilige Strukturebene (Klammern). XML ist sehr gut geeignet, um Daten strukturiert darzustellen. Was fehlt, ist der Bezug zum gedruckten Dokument. Dieser kann durch zusätzliche Formatierungsinformationen hergestellt werden. Sei es durch CSS (Cascading Style Sheets), einer Stilsprache, die das Aussehen von HTML-Dokumenten definiert; oder durch XSL-FO (s.u.), die beschreibt, wie Text, Bilder, Linien und andere grafische Elemente auf einer Seite angeordnet werden, vor allem von Dokumenten, die gedruckt oder archiviert werden sollen.

HTML basiert auf XML und beschreibt spezifische Tags zur Auszeichnung von Dokumenten. Dabei versucht HTML, die Semantik abzudecken und das Layout durch CSS festzulegen. Hierbei sind in HTML5 noch einige deutliche Verbesserungen gegenüber HTML4 vorgenommen worden. Aus Kompatibilitätsgründen akzeptieren Browser aber bis heute noch invalides XML, so dass HTML nicht von allen XML-Tools unterstützt wird. Ein weiterer Vorteil: XML ist leicht zu generieren, zu verarbeiten und beliebig erweiterbar. Zudem gibt es inzwischen eine Vielzahl etablierter XML-Tools.

Verbreitung:

Angesichts dieser Vorteile hat XML die Sprache SGML (Standard Generalized Markup Language) als ISO-Standard für die Dokumentenrepräsentation nahezu vollständig abgelöst. XML ist in vielen Bereichen des Lebens zu finden, u.a. als

- RSS-Feed
- Login beim E-Mail-Account (via SAML)
- XHTML-basierte Webseiten
- PowerPoint-Präsentationen
- Leistungsabrechnungen
- Konfigurationsdateien

Branchen mit hoher XML-Verarbeitung:

- Gesundheits- und Sozialwesen
- öffentliche Verwaltungen
- Finanzdienstleister
- ITK
- Energieversorgung

XSL-FO (Extensible Stylesheet Language – Formatting Objects)

XSL-FO ist eine auf XML basierende Auszeichnungssprache, die beschreibt, wie Text, Bilder, Linien und andere grafische Elemente auf einer Seite angeordnet werden. XSL-FO enthält unter anderem folgende Elemente und Attribute (Auswahl):

- Regionen, Ränder und Bereiche einer Seite
- Breite, Höhe und Abfolge von Seiten
- Seitenverwaltung
- Rahmen, Abstände, Blöcke
- Absätze, Listen und Tabellen
- Textformatierung wie Satzformate Zeilenumbrüche und Trennung
- Linien, Bilder und andere Objekte

XSL-FO wurde 2001 vom W3C als Standardsprache für die Umwandlung von XML-Dokumenten in Druckformate festgelegt. Mittlerweile existieren dafür zahlreiche Formatierungslösungen.

Vorteile:

- wird vor allem für die Formatierung von großen Dokumenten mit vielen Seiten eingesetzt
- sehr gut geeignet für die Strukturierung von Texten/Dokumenten und die Definition von Kategorien wie Adresse, Name, Vorname, Geschlecht etc.
- Verknüpfung von Daten, beispielsweise Adress- und Rechnungsdaten, und deren Anordnung im jeweiligen Format möglich
- Aufgrund ihrer XML-Basis können XSL-FO-Lösungen in bestehende XML-Verarbeitungsprozesse eingebunden werden. Dies wirkt sich vorteilhaft auf automatisch erzeugte Dokumente aus.
- Qualitätskontrolle möglich durch Validierung mit XML-Schema oder DTD
- durch den eigenen Namensraum von XSL-FO lassen sich sehr einfach eigene Erweiterungen realisieren
- Kombinationen mit anderen Auszeichnungssprachen wie SVG möglich

XMP (Extensible Metadata Platform)

- Standard für die Einbettung von Metadaten in digitale Dateien
- von Adobe im Jahr 2001 veröffentlicht und erstmals in den Acrobat Reader 5 integriert
- Februar 2012: Veröffentlichung des Kernteils der XMP-Spezifikation als ISO-Standard ISO 16684-1

XMP basiert auf offenen Standards und bettet die vom W3C veröffentlichte formale Sprache RDF (Resource Description Framework) in Binärdaten ein. Damit sollen die Metadaten in verschiedenen Applikationen nach einem einheitlichen Schema so integriert werden, dass die Dateien auch weiterhin von anderen Programmen gelesen werden können. Das Format wird von allen Adobe-Produkten, der Software anderer Hersteller sowie Anbietern von Redaktionssystemen unterstützt.

XMP definiert unter anderem:

- die Sprache des Dokuments (eines der wichtigsten Merkmale; besonders wichtig für Menschen mit Sehbehinderung/Vorlesen des Dokuments mittels Screen-Reader in der korrekten Sprache)
- das Erstellungsdatum
- Autor/Name der Firma (Woher kommt das Dokument?)
- Stichworte/Keywords

RDF (Resource Description Framework)

RDF ist eine technische Herangehensweise im Internet zur Beschreibung von Ressourcen (Subjekt, Prädikat, Objekt) und ihrem Verhältnis zueinander. Ursprünglich wurde RDF vom W3C als Standard zur Definition von Metadaten konzipiert. Mittlerweile gilt RDF als ein grundlegender Baustein des „semantischen Webs“. RDF ähnelt den klassischen Methoden zur Modellierung von Konzepten wie UML-Klassendiagramme und Entity-Relationship-Modell.

Über Standardisierungsbestrebungen wurden häufig benutze Aussagen über ein Objekt zu so genannten Ontologien zusammengefasst, die über einen Namensraum URI (Universal Resource Identifier) identifiziert werden. Dies ermöglicht unter anderem Programmen, Daten für den Menschen sinnvoll darzustellen.

Ontologie

Eine Ontologie ist eine Sammlung von Begriffen, mit denen man Metadaten definiert, unter anderem Titel, Autor, Thema, Beschreibung, Datum, Sprache, Ort der Aufnahme und Kameratyp (bei Bildern/Fotos). Gängige Ontologien sind Dublin Core, IPTC, Exif.

ZUGFeRD

Bei ZUGFeRD handelt es sich um ein einheitliches Format für elektronische Rechnungen, entwickelt vom „Forum elektronische Rechnung Deutschland“ (FeRD). Es bietet die Kombination der visuellen Repräsentation eines Dokuments mit seinen Rohdaten in einer einzigen PDF/A-3-Datei zur Vermeidung manueller Eingriffe in der automatischen Verarbeitungskette.