Das PDF – es geht nicht ohne.

Dokumentenformate, ISO-Standards, Langzeitarchivierung

Nicht alle Dokumentenformate haben es zum Standard geschafft und lange nicht jedes hält, was es verspricht. Was kann oder muss man von einem Format erwarten? Was sind die wichtigsten Qualitätsmerkmale eines Formates und wie findet man heraus, ob ein Format „gut" ist? Wo lauern die Gefahren und wie umgeht man diese? Die Formatwahl spielt vor allem im Bereich Langzeitarchivierung eine große Rolle. Eine jahrelange Archivierung von Dokumenten in einem ungeeigneten Format oder in schlechter Qualität kann zu ernstzunehmenden Konsequenzen führen.

PDF und die ISO-Normen

Ein Format sticht dabei besonders heraus: PDF, das native Format von Adobe Acrobat. Es war von Anfang an für den Austausch von Dokumenten, unabhängig von Plattform und Software, konzipiert. Es wird von fast jedem Softwarehersteller implementiert, der elektronische Dokumente erzeugt, und hat sich seit 21 Jahren bewährt. Allerdings hat die astronomische Verbreitung von PDF die Angst vor der Abhängigkeit von Adobe geschürt. Um diesem Konflikt zu entgehen, haben Adobe sowie einige Anwender und Industrieunternehmungen sich dafür eingesetzt, dass PDF zum Industriestandard wird und durch die Gremien von ISO weiterentwickelt wird. Das erste Resultat dieser Initiative ist die Publikation des Standards ISO 32000-1 im Jahr 2008. Er basiert auf der Version PDF 1.7 von Adobe. Zurzeit arbeitet ISO an der Version PDF 2.0.

Der groß gewachsene Funktionsumfang von PDF ist nicht für jeden Anwendungsbereich geeignet und auch nicht notwendig. Deshalb hat ISO eine Anzahl von Unternormen entwickelt, welche auf dem PDF-Standard aufgebaut und speziell auf die wichtigsten Anwendungen zugeschnitten sind (siehe Bild 1).

Die Normen PDF/X, PDF/A, PDF/E, PDF/VT und PDF/UA sind nicht eigene Dateiformate. Sie basieren alle auf der übergeordne-

www.pdf-tools.com

Dr. Hans Bärfuss ist Gründer und Geschäftsführer der **PDF Tools AG**, einer international erfolgreichen Softwareentwicklungs- und Vertriebsgesellschaft. Er ist Delegierter der Schweizerischen Normenvereinigung (SNV) bei der ISO und hilft bei der Standardisierung von Dateiformaten und digitalen Signaturen mit. Er ist einer der Initiationen und Gründer des PDF/A Competence Centers, eines Verbands zur Bekanntmachung und Förderung des ISO Standards PDF und PDF/A, und Chairman des Swiss Chapter und hält zahlreiche Fachvorträge auf Konferenzen und Seminaren und publiziert Fachartikel zum Thema digitale Dokumente.

Norm	Zweck
PDF/X	Austausch von Druckvorlagen: Das Druckergebnis soll vorhergesagt werden können.
PDF/A	Langfristige, elektronische Archivierung: siehe Text
PDF/E	Ingenieurwesen: interaktive 3D-Modelle für Konstruktionszeichnungen und Produktions-unterlagen
PDF/VT	Variabler Datendruck und Transaktionsdruck: großes Volumen mit Caching und Streaming
PDF/UA	Universal Access: Barrierefreiheit und Bedienungshilfen, wie Screen-Reader

Bild 1: Die wichtigsten Unternormen von PDF

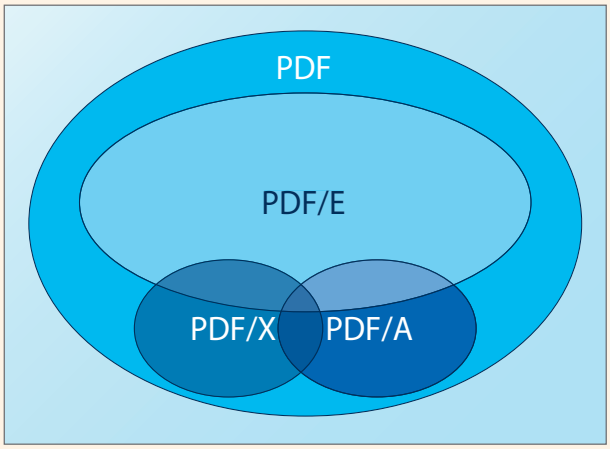

Bild 2: Die Unternormen von PDF definieren eine Untermenge des Funktionsumfangs

ten Norm PDF 1.7 und definieren die für den jeweiligen Anwendungszweck geeigneten Gebote und Verbote. Diese Regeln schränken den Funktionsumfang von PDF 1.7 entsprechend ein und bilden eine definierte Untermenge (siehe Bild 2).

PDF für die langfristige, elektronische Archivierung

Obwohl PDF vor allem für den Austausch von Dokumenten, unabhängig von Betriebssystem sowie Erzeuger- und Betrachter-Software, entwickelt wurde, ist das Format nicht ohne weiteres für die Archivierung geeignet. Dafür wurde PDF/A entwickelt.

Mit PDF/A wird eine Reihe von Normen bezeichnet, welche die Eigenschaften von PDF-Dateien für die langfristige elektronische Archivierung von Dokumenten beschreiben. Die Normen bezwecken, dass die archivierten Dokumente langfristig zugänglich bleiben, selbstbeschreibend und eindeutig sind, durch Metadaten beschrieben werden und durchsuchbar sind. Aus diesen Gründen ist die Verschlüsselung verboten, müssen Schriften und Farbprofile eingebettet sein, alternative und dynamische Inhalte entfernt und Bezüge auf externe Inhalte aufgelöst werden.

Es ist grundsätzlich auch möglich, dass eine PDF-Datei mehreren Unternormen gleichzeitig entspricht. Dies ist insbesondere für die Archivierung von PDF/X und PDF/VT konformen Dateien wichtig, welche gleichzeitig der PDF/A Norm entsprechen müssen. Der Funktionsumfang einer Datei, welche die Konformität mit mehreren Normen beansprucht, schränkt sich entsprechend auf die Schnittmenge der einzelnen Normen ein.

Der Wettbewerb der Formate

PDF ist ein sehr leistungsfähiges Format, jedoch bei weitem kein Alleskönner. Es gibt eine große Zahl von Formaten, welche für spezifische Anwendungszwecke optimiert sind (siehe Bild 5).

Wenn es um die Beschreibung der Vor- und Nachteile von Dateiformaten geht, richten viele Autoren den Blick spontan auf PDF und veröffentlichen mehr oder weniger fundierte Vergleiche zwischen diesen Formaten und PDF. Damit der Vergleich einigermaßen fair ausfällt, darf man jedoch den Anwendungszweck nicht aus den Augen verlieren. Nur wo es Überschneidungen in den Anwendungsbereichen gibt, sind Vergleiche sinnvoll. Betrachten wir nun die am häufigsten diskutierten Bereiche.

Posteingang

Geht es um die Bilder gescannter Dokumente, ist TIFF weder

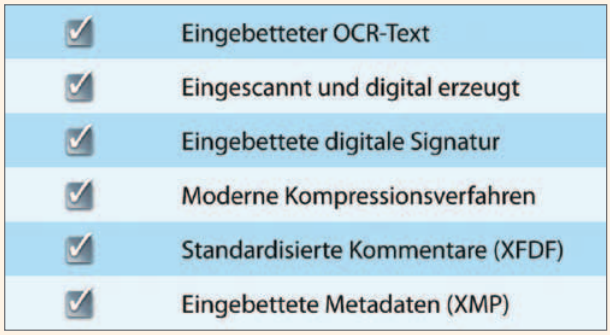

Bild 3: Kriterien für die Migration von TIFF nach PDF/A

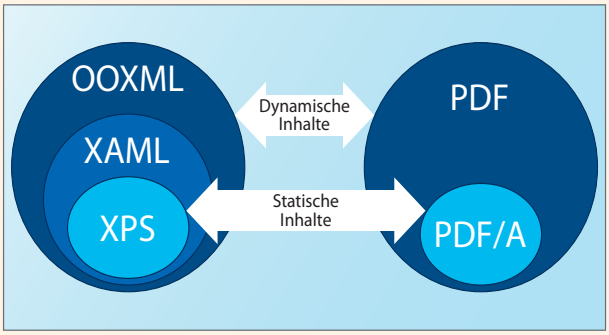

Bild 4: Vergleich Office-Formate mit PDF

besser noch schlechter als PDF geeignet. Erst wenn zusätzliche Anforderungen gefragt sind, welche über die reine Seitendarstellung hinausgehen, zeigen sich die Vorteile von PDF. TIFF-Archive sollten auch nicht blind in PDF-Archive migriert werden, wenn kein Mehrwert geschaffen wird, der den Aufwand rechtfertigt. Die Kriterien für eine Migration von TIFF nach PDF sind in Bild 3 aufgeführt.

Arbeitsdokumente

Die Office-Formate werden am häufigsten für Arbeitsdokumente gebraucht. OOXML ist das native Format von Word, Excel und PowerPoint. Ein fairer Vergleich zwischen OOXML und PDF ist nicht möglich. OOXML ist für die Erstellung und das Editieren entworfen, beinhaltet Strukturinformationen und stellt durchsuchbaren Text bereit (Copy/Paste). PDF als Autorenformat oder sogar als geeignet für das Editieren zu bezeichnen, ist ziemlich gewagt. Dafür punktet PDF als Format mit festem Layout und für die Archivierung, wofür OOXML definitiv nicht geeignet ist, auch wenn dies einige Anwender glauben möchten.

Vergleichbar hingegen sind die jeweiligen Untermengen XPS und PDF/A, welche viele Eigenschaften, wie statische Inhalte, gemeinsam haben. Das Motiv für die Entwicklung von XPS war, dass zwischen dem Dokument und dem Betriebssystem keine Formatbrüche mehr existieren sollen. So ist das XPS-Graphikmodell das gleiche, wie das des neuen Graphiksubsystems WPF von Vista. Genauer gesagt: XPS-Elemente sind eine Untermenge von XAML, der in WPF enthaltenen Beschreibungssprache für Dokumente und Benutzerschnittstellen (siehe Bild 4).

Allerdings zeigt sich gerade in diesem Fall, wie wichtig die Verbreitung eines Formats ist. PDF/A ist unterdessen der De-facto-Standard in der Archivierung. Im Gegensatz dazu ist XPS nahezu unbekannt. Für ODF, das von Microsoft unabhängige Office-Format, gelten ähnliche Schlussfolgerungen.

Postausgang

Druckdatenströme im PostScript-, PCL- und AFP-Format werden manchmal direkt archiviert, öfters jedoch in TIFF konvertiert (COLD). In den letzten Jahren überlegt man sich die „Vertiffung", wie sie manchmal genannt wird, durch die Umwandlung in PDF/A zu ersetzen. Dabei spielen die Kriterien, wie Dateigröße (Zwang zur Einbettung von Schriften), Aufwand für die Umwandlung, Qualität des Resultats usw. eine wichtige Rolle. Für PostScript und PCL gibt es keinen Archivierungstandard. Für AFP hingegen arbeitet man an einer AFP/A-Norm als Alternative zu PDF/A.

Diese Überlegungen spielen vor allem dann eine Rolle, wenn man für den Postausgang eine eigene Archivierungslösung vorsieht. Gibt man einer unternehmensweiten Archivierungslösung den Vorzug, landet man schnell wieder bei PDF/A.

Dateiformat	Wurde entworfen für
TIFF	Austauschformat für Rasterbilder (Scanning, Archivierung)
PostScript/PCL	Seitenbeschreibungssprache für den Druck
PDF	Austausch von Dokumenten mit festem Layout und interaktiven Elementen
AFP	Transaktionsdruck mit variablen Daten
Office Formate	Proprietäres Format für Arbeitsdokumente in der Unternehmung
OOXML	Open Office-Format auf XML-Basis von Microsoft
ODF	Open Office-Format in Konkurrenz zu Microsoft
XML	Speicherung von objektorientierten Daten mit Bezug zu einem Schema
XMP	Erweiterbares Metadatenformat auf XML-Basis
XPS	Auf XML basierende Seitenbeschreibungssprache von Microsoft
EPUB	Format für die Publikation von Dokumenten im Internet
PRC	Speicherung von 3D-Daten und Attributen
Multimedia	Zahlreiche Formate für das Audio- und Videostreaming

Bild 5: Dateiformate und Verwendungszweck

3D-Daten für das Ingenieurwesen

Für 3D-Daten, welche interaktiv manipuliert werden können und mit Beschreibungen angereichert sind (z.B. für Stücklisten), gibt es eine ISO-Norm mit dem Namen PRC (Product Representation Compact). PRC lässt sich in PDF einbetten und ist ein wichtiges Merkmal der PDF/E-Norm. PDF/E-2 ist so gestaltet, dass konforme Dateien direkt archiviert werden können.

Metadaten

Für Metadaten gibt es zahlreiche proprietäre Formate. Metadaten werden oft auch direkt im Archivsystem geführt. Es empfiehlt sich aber dringend, auch für diesen Zweck ein Standardformat zu verwenden. ISO bietet dafür XMP (Extensible Metadata Platform) an, welches auf XML basiert. XMP kann in fast alle Bildformate wie JPEG, TIFF usw. eingebettet werden und ist notwendiger Bestandteil von PDF/A. In PDF/A-3 lassen sich auch direkt XML-Daten, beispielsweise für die elektronische Rechnungsstellung, einbetten (ZUGFeRD).

Multimedia

Audio- und Videodateien werden selbstverständlich auch archiviert, sowohl einzeln als auch als eingebettete Datenströme in PDF und anderen Formaten. Allerdings gibt es für diese Formate noch keine allgemein anerkannten Standards. Dies wird sich hoffentlich in Zukunft ändern. Dabei spielt EPUB, ein Format für die elektronische Publikation, eine treibende Rolle.

Archivierung – ist PDF/A die Lösung?

Solange man sich in der Welt des elektronischen Dokuments befindet, kann PDF/A bestimmt die meisten Anforderungen an ein Dateiformat erfüllen. Allerdings gilt dies bei weitem nicht für jedes mögliche Archivgut. Vor allem für interaktive Kommunikationsmittel wie Webseiten, Programme und Multimedia-Inhalte braucht es neue Normen. Den Experten von ISO wird also die Arbeit vorerst kaum ausgehen.