PDF-Dateien in der Archivierung

Auf zum neuen Archiv

Als „PDF/A“ hat das Dateiformat PDF den Schritt vom de facto zum de jure-Standard gemacht. Für die Archivierung von Daten steht nun ein transparentes und vielseitiges Dateiformat bereit.

Carsten Heiermann

Schon lange zählt PDF wegen seiner Layouttreue zu den beliebten Datei-Formaten. Der Dokumentencharakter von beispielsweise Briefen oder anderen Korrespondenzen bleibt erhalten. Die Größe der Dateien ist im Vergleich zu anderen Formaten klein, und jeder, der den kostenlosen Adobe Reader auf seinem Rechner installiert hat, kann PDFs lesen. Was bisher fehlte, war die Sicherheit, dass PDF-Dateien auch langfristig lesbar bleiben. Dazu erlaubte PDF bisher zu viel, wie zum Beispiel das Verweisen auf externe Quellen oder den Einsatz von eigenen Schriften, die sich unabhängig vom eigentlichen Dokument im Laufe der Jahre ändern und so die unveränderte Darstellung nicht garantieren können. Es war notwendig, ein standardisiertes PDF-Format zu entwickeln. Ein Format, das sich zur Darstellung und ‚Langzeitarchivierung' eignet. Denn nur dann ist beispielsweise die gesetzeskonforme und revisionssichere Vorhaltung der Dokumente möglich. Die Standardisierung erfolgte auf Basis fundamentaler Ansprüche an ein Dateiformat für die Langzeitarchivierung. Dazu zählen:

(1) Geräte-, Software- und Versionsunabhängigkeit, sodass die Inhalte immer gleich dargestellt werden.

(2) „Self Contained“, beinhaltet alle Komponenten, die zur Darstellung nötig sind, in der Datei.

(3) „Self Documented“, die Dateien beschreiben sich inhaltlich und dokumentieren sich selbst über wiederum standardisierte Metadaten.

(4) Transparenz, eine PDF/A-kompatible Datei ist mit einfachen Mitteln analysierbar.

PDF/A wird ISO-Standard

Um diesen Standard zu entwickeln, hat die AIIM (Enterprise Content Management Association) und die NPES (Association for Suppliers of Printing, Publishing, and Converting Technologies) im Oktober 2002 eine Arbeitsgruppe ins Leben gerufen. Diese Arbeitsgruppe setzt sich aus Vertretern der Forschung, der öffentlichen Hand und der Wirtschaft zusammen. Adobe, Appligent, EMC, Global Graphics zählen genauso dazu wie die Harvard Universität, IBM, das nationale Staatsarchiv der USA NARA, die Library of Congress, Merck & Co., das Patent- und Markenamt der USA, PDF Sages und US-Gerichte. Außerdem gehören Vertreter nationaler Normungsinstitute zum Gremium. Im Herbst vergangenen Jahres war es so weit: Die International Organization of Standardization (ISO) hat PDF/A, das A steht sinnigerweise für Archive, als Standard für die

Langzeitarchivierung anerkannt, und es wird vermieden, dass Unternehmen für die Archivierung eigene Unterdialekte von PDF für die Langzeitarchivierung erzeugen. Die notwendigen Limitierungen von PDF müssen also nicht in jedem Unternehmen immer wieder neu als Unternehmensrichtlinie definiert werden. Gleichzeitig ist sichergestellt, dass Mitarbeiter die Dokumente auch nach Jahren noch lesen können, wenn sie im PDF/A-Format abgespeichert werden.

Vorteil der Standards

Grundsätzlich haben standardisierte Formate den Vorteil, dass Anwender nicht an bestimmte Hersteller gebunden sind und sich somit in eine gewisse Abhängigkeit begeben. Der Grund für die Entwicklung eines neuen Standards für die Langzeitarchivierung von Dokumenten liegt in den Schwächen der bisherigen Standardarchivformate. Dies waren nach dem IT-Grundschutzbuch des Bundesamtes für Sicherheit in der Informationstechnik (www.bsi.de/gshb/deutsch/m/m04170.html) die Bildformate TIFF und JPEG. Auch diese standardisierten Formate sind im Gegensatz zu nativen Formaten herstellerunabhängig lesbar und werden von zahlreichen Programmen unterstützt. Doch TIFF und JPEG haben Nachteile. Zum einen verändern sie das originäre Aussehen der Datei. Davon betroffen sind beispielsweise Farbinformationen, die bei der Umwandlung in das weit verbreitete TIFF G4 verloren gehen. JPEG erzeugt entweder schlechte Darstellungsqualitäten oder sehr große Dateien. Aber auch die in der Richtlinie nicht genannte, gleichwohl von manchen Experten empfohlene Ablage im Originalformat (Excel- oder CAD-Dateien) scheint keine Lösung für die Langzeitarchivierung. Eine immer gleiche, eindeutige Darstellung des Inhaltes – noch dazu versionsunabhängig – ist damit kaum möglich. Ein Beispiel dafür sind die vielen Darstellungsmöglichkeiten und Ansichten einer einzigen Microsoft-Project-Datei oder die Versionierungen in Word.

Darüber hinaus sind Bildformate nicht volltextrecherchefähig, da sie – wie der Name schon sagt – Dokumente als Bild ablegen und kein interoperabler Standard für die Einbettung solcher OCR-Informationen oder Metadaten existiert. Die Sinnhaftigkeit von Buchstaben geht durch das „Rastern“ oder „Verpixeln“, also die Umwandlung von strukturierten Dokumenten in reine Bilder, verloren. Die TIFF-Dateien werden zwar schnell und ohne groß nachzudenken eingesetzt, aber TIFF ist kein internationaler Standard, sondern eine im Recht von Adobe stehende Herstellerfestlegung. Zu guter Letzt sind TIFF- und JPEG-Dateien mit mehr als 15 Jahren „uralt“ im Verhältnis zur allgemeinen IT-Innovationsgeschwindigkeit – was als solches kein Grund ist, das Format abzulehnen.

Das bessere PDF hat ein „A“

Neben den bekannten Formaten wird auch das PDF-Format im Maßnahmenkatalog des IT-Grundschutzbuches des Bundesamtes für Sicherheit in der Informationstechnik diskutiert. PDF eignet sich „primär für die Archivierung von Dokumenten, bei denen eine Abbildung in Papierform vorgesehen ist oder die den Charakter von Briefen und Geschäftsdokumenten haben.“ Aber, so der Maßnahmenkatalog weiter, „PDF ist nicht standardisiert. Wenn es als Datenformat zur elektronischen Archivierung verwendet werden soll, sollte das Datenformat PDF separat dokumentiert werden.“

Der PDF/A-Standard (ISO 19005-1) schreibt dagegen detailliert vor, welche Inhalte erlaubt sind und welche nicht. Er stellt quasi eine eingegrenzte Variante von PDF dar, ein standardisiertes Profil zur Verwendung von PDF in der Archivierung. Dadurch soll eine langfristige, unveränderbare und universelle Lesbarkeit der Dokumente garantiert sein – unabhängig davon, mit welcher Anwendungssoftware und auf welchem Betriebssystem sie ursprünglich erstellt wurden. So dürfen im Vergleich zum „normalen“ PDF keine Fremdabhängigkeiten oder Referenzen zu externen Quellen integriert sein, Schriften müssen komplett eingebunden sein.

„-1“ steht für den ersten Teil der Standardisierung von PDF/A. Dabei wurde auf Adobes PDF-Version 1.4 zurückgegriffen. Experten wissen, dass dies eine Einschränkung darstellt, denn für die Archivierung sinnvolle Features wie die eingebettete elektronische Signatur wurden erst später

in PDF-Version 1.6 eingeführt. Nicht zuletzt deshalb arbeitet die ISO-Kommission derzeit an Teil zwei des Standards (19005-2). Er wird auf PDF 1.6 basieren. Eine andere Notation ist: PDF/A-1a und PDF/A-1b. Dabei handelt es sich um zwei Level, „Level a" heißt volle Compliance mit den stärksten Einschränkungen und insbesondere der Pflicht, neben den Inhalten der Dateien auch deren Aufbau noch in den Metadaten dokumentieren zu müssen. Für viele Anwendungsfälle reicht die „Level b"-Compliance aber vollkommen aus.

Lesbare Metadaten

Neben den Dokumentinhalten spielen Metadaten eine große Rolle, die „die Daten über den Daten" sind. Mit dem PDF/A-Format können sie, wie übrigens bei den meisten anderen Format-Typen auch, genutzt werden. In einem Papierarchiv sind die Metadaten Indizes, Dateilisten, Register, Ordnerrücken oder andere Hilfen, welche die Suche eines Dokuments erleichtern sollen. Die Metadaten elektronischer Dokumente werden über ein automatisiertes System oder den Autor der Informationen erfasst. Die Metadatenschemata im PDF-Format sind sowohl von Anwendern als auch von IT-Systemen lesbar. Das zugrunde liegende Bindeglied XML erfordert lediglich eine Definition aller Namensbereiche wie Ersteller, Titel und Beschreibung. PDF-Dateien enthalten Metadaten, die die Eigenschaften eines Dokumentes beschreiben, sind aber nicht darauf beschränkt.

Alle Änderungen, die im Dialogfeld „Dokumentzusammenfassung" in Adobe Acrobat vorgenommen werden, spiegeln sich in den Metadaten wider. Weil Metadaten im XML-Format vorliegen, können sie mit Produkten von Fremdherstellern erweitert und abgeändert werden. Zusätzlich sind in PDF/A noch Metadaten gemäß XMP-Standard enthalten. Mit diesem von Adobe definierten, herstellerspezifischen Standard lassen sich Informationen über Dateien lesen und bearbeiten, ohne dass die Datei dazu geöffnet werden muss. Solche Metadaten werden benutzt, um Inhalt und Struktur der PDF-Datei zu beschreiben. Unter Struktur kann man sich vorstellen, dass es eine nach XMP standardisierte „Bauanleitung" gibt, in der verzeichnet ist, wie aus den in der Datei befindlichen Objekten (Bilder, Schriften, Texte u. a.) der eigentliche Inhalt des Dokumentes zusammengesetzt wird. Damit schaffen diese Metadaten Versionsunabhängigkeit. Zusätzlichen dienen sie der Selbst-Dokumentation und erlauben auch eine sehr schnelle Analyse des Dateiinhaltes, wodurch sie dem Grundsatz der Transparenz von PDF/A dienen.

Wie kann der Standard genutzt werden?

Ein entscheidender Vorteil von PDF/A ist die universelle Einsatzfähigkeit. Sowohl für gescannte Dokumente als auch für vektorisierte Dateien oder Office-Dokumente, CAD-Zeichnungen kann einheitlich PDF/A genutzt werden. Das ging bisher nur mit PDF, mit den oben genannten Schwierigkeiten und Bedenken. Es ist nicht notwendig, Word-Dateien zunächst in TIFF gerastert zu speichern. Der Anwender kann sie direkt ins PDF/A-Format konvertieren und hat sie immer noch vektorisiert und trotzdem revisionssicher abgelegt. Die Ablage vektorisierter Daten erspart gegenüber in Bilddaten umgewandelten Dateien Aufwände für Schrifterkennung (OCR) und erlaubt ein originalgetreues Weiterverarbeiten von Textteilen dieser Dateien. Aufgrund dieser universellen Einsatzfähigkeit des PDF/A-Formates sowohl für Raster- als auch Vektordateien hat PDF/A ein gutes Potenzial, sich als einheitliches Archivformat für alle Materialien durchzusetzen.

PDF/-A-kompatible Dokumente können – genauso wie „normale" PDF-Dateien – mit einem Viewer gelesen werden. Unternehmen, die sich mit dem Standard en Detail auseinandersetzen wollen, können von der ISO (www.iso.org) oder der AIIM (www.aiim.org/bookstore) die Dokumentation erwerben, aus der hervorgeht, welche Reglementierungen PDF/A vorschreibt.

Wie entsteht ein PDF/A?

Viel wichtiger ist aber: Wie erzeugt man ein PDF/A-kompatibles Dokument? Natürlich zählt Adobe mit Acrobat 7.0 selbst zu den ersten Anbietern, die den PDF/A-Entwurf unterstützen. Doch bei der Betrachtung, welche Software die Konvertierung in PDF/A ermöglichz, trennt sich schnell die Spreu vom Weizen: Echte Lösungen sind noch Mangelware. Es scheint jedoch, dass viele Hersteller PDF/A in Zukunft unterstützen werden, da die Nachfrage seitens der Archivbetreiber aufgrund der Vorteile des neuen Formates sehr hoch ist. Wie kann der Anwender also beispielsweise aus einer Office-Anwendung oder einem anderen Programm eine Datei generieren, die dem PDF/A-Standard entspricht? Jede Software muss dazu die Originalformate in allen Versionen lesen und interpretieren können, was in vielen Fällen aber rechtlich bedenklich sein kann. Es wird also die Notwendigkeit bestehen, PDF/A unter Nutzung der zum jeweiligen Dateiformat gehörenden Anwendungen zu erzeugen. Virtuelle Druckertreiberlösungen werden daher wohl die häufigste Form von Tools zur PDF/A-Erzeugung sein.

Nicht zu vergessen sind Tools zur Verifikation der PDF/A-Kompatibilität. Auch wenn der PDF/A-Standard ein vergleichsweise kurzes Dokument ist, ist er mit den häufigen Verweisen auf die sehr voluminöse PDF-Spezifikation doch ein kompliziertes Regelwerk. An der Schnittstelle zum Archivsystem sollte also ein Verifikationsmodul darüber wachen, dass die zu archivierende PDF-Datei wirklich dem gewünschten Level von PDF/A entspricht. Sind zum Beispiel alle Schriften eingebettet? Sind alle Metadaten vorhanden? Beinhaltet die Datei keine unzulässigen Komponenten? Das alles kann nur ein Verifikationsmodul feststellen und entweder die Datei für das Archiv zulassen oder deren „Fehler" dokumentieren. Viele Hersteller haben solche Tools angekündigt, eine offizielle Stelle zur Überprüfung von PDF/A-Dateien oder gar zur Zertifizierung von Verifikationssoftware gibt es aber nicht.

Durch die formale Anerkennung der ISO wird PDF vom De-facto-Standard, also aufgrund seiner häufigen Verwendung und Akzeptanz zu einem De-jure-Standard gehoben. Gleichzeitig werden die Probleme von PDF für die Langzeitarchivierung durch den PDF/A-Standard behoben. Trotz der bisher nur wenigen am Markt verfügbaren Lösungen gehen DMS-Experten davon aus, dass sich PDF/A als Standardformat für die Langzeitarchivierung durchsetzen wird. Seine Vorteile überwiegen im Vergleich zu den anderen (Bild-)Formaten. Es ist das vielseitigere Format, das durch die Volltextfähigkeit und die oben beschriebenen Merkmale nützliche Optionen beinhaltet. –

Dipl. Ing. Carsten Heiermann

Der studierte Elektrotechniker, Jahrgang 1968, war treibende Kraft bei der Fusion der LuraTech mit der Algo Vision 2001. Seit 2004 ist er einer der Gesellschafter der LuraTech GmbH und Initiator des PDF/A-Kompetenzzentrums.