White Paper

PDF/A – Ein neuer Standard für die Langzeit-Archivierung

PDF Tools AG

- → **Warum die PDF/A-Initiative?**
- → **Was ist die PDF/A-Norm?**
- → **Was ist PDF/A-1a, PDF/A-1b, PDF/A2?**
- → **Wie soll die PDF/A-Norm eingesetzt werden?**
- → **Liefert PDF/A die Antwort zur Langzeitarchivierung?**

pdf-tools.com
Premium PDF Technology

Version: 2.0

Datum: 22. Januar, 2007

Inhaltsverzeichnis

Einführung

PDF/A - A new Standard for Long-Term Archiving

Hintergrund

Am 28. September 2005 hat die Internationale Organisation für Standardisierungen (ISO) einem neuen Standard für die Regelung der Archivierung elektronischer Dokumente zugestimmt:

ISO-19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).

Dieser Standard ist das Ergebnis einer über dreijährigen Sitzungsarbeit von Vertretern aus weltweit ansässigen Unternehmungen und anderen Organisationen.

PDF/A wurde als internationaler Standard im September 2005 genehmigt.

Mehrere Organisationen, Hersteller und Anwender waren daran beteiligt.

Im Mai 2002 lancierte die AIIM (Association for Information and Image Management), die NPES (National Printing Equipment Association) und die Verwaltung der US Gerichte in den USA eine Initiative für die Schaffung eines Standardformats für elektronisch archivierte Dokumente. Das kick-off Meeting wurde im Oktober 2002 durchgeführt. Daran nahmen PDF Hersteller wie Adobe Systems, Library of Congress, Surety Inc., Quality Associates Inc., Appligent, Merck, EMC, PDF Sages, and NARA (National Archives & Records Administration) teil. Auch Xerox, Honeywell, EDS, and Glaxo Smith Kline stiessen später dazu, um nur einige zu nennen.

Die Urheber der Initiative bereiteten einen ersten Entwurf vor und reichten ihr Projekt bei der ISO für die Registrierung als internationaler Standard ein. Die ISO wies das Projekt dem technischen Komitee TC 171 (Document Management Applications) zu. Das TC 171 besteht aus Vertretern von 13 Mitgliedländern (die je eine Stimme haben) sowie Beobachter weiterer 21 Länder. Nach zahlreichen Durchläufen und Verbesserungen wurde der Standard im September 2005 angenommen.

Warum die PDF/A Initiative?

Archivierungsformate variieren von Land zu Land. Traditionelle Archivierungsmethoden (Papier, Mikrofilm, Mikrofiche) garantieren zwar die Reproduzierbarkeit, entsprechen aber nicht mehr dem neusten Stand der Technik. Grosse Dokumente können nicht rasch um den Globus geschickt werden und es ist ausserordentlich schwierig, archivierte Dokumente nach bestimmten Inhalten zu durchsuchen. Um einen ersten Schritt Richtung elektronische Archive zu gehen, haben viele Organisationen TIFF-Archive eingerichtet. Auch TIFF garantiert die Reproduzierbarkeit auf lange Sicht und ist ein etabliertes Format. TIFF kann nun zwar schnell und einfach in weltweit vernetzten Unternehmungen übermittelt werden, die Suche hingegen ist nach wie vor schwierig.

Neue Technologien haben die elektronische Archivierung von Dokumenten ermöglicht.

PDF weist mehere Vorteile gegenüber TIFF auf.

Man begann nun den Blick auf PDF zu richten. Eine Reihe von Gründen machen PDF gegenüber TIFF attraktiver:

- PDF speichert strukturierte Objekte (wie Texte, Vektorgraphiken, Rasterbilder), welche das effiziente Suchen im ganzen Archiv unterstützten. TIFF

hingegen ist ein Rasterformat und muss vorgängig mit einer OCR-Maschine bearbeitet werden, um eine Volltextsuche zu ermöglichen.

- PDF Dateien sind kompakter und benötigen oft nur einen Bruchteil des Speicherplatzes einer entsprechenden TIFF-Datei oft auch noch mit besserer Qualität. Die geringere Dateigrösse ist vor allem für den elektronischen Datenaustausch von Vorteil (FTP, Email Anhänge usw.).
- Metadaten wie Titel, Autor, Erstellungs- und Modifikationsdatum, Inhalt, Schlagworte usw. können direkt in die PDF-Dokumente eingebettet werden. Dadurch lassen sie sich ohne menschliches Zutun automatisch klassifizieren.
- Die Seiteninhalte in einem PDF-Dokument sind in der Regel geräteunabhängig, d.h. von der Rasterauflösung, dem Farbsystem, usw. unabhängig. Die Seiten werden erst bei der Reproduktion auf ein Raster abgebildet (Rendering-Prozess). Damit profitieren PDF-Dokumente vom technologischen Fortschritt der Ausgabegeräte wie Drucker, Bildschirme usw., auch nach Jahren noch.

Die Schöpferin des PDF de facto Standards, Adobe Systems, hat im Laufe der letzten zwölf Jahre sieben neue Versionen ihres "PDF Reference Manual" publiziert. Jede neue Version hat das Format um zahlreiche neue Eigenschaften angereichert und einige alte verändert. Es war deshalb notwendig, aufbauend auf Adobe’s proprietärer PDF Spezifikation einen stabilen, international akzeptierten Standard für die Langzeit-Archivierung zu schaffen. Das Resultat: PDF/A.

Der PDF/A Standard

Das Ziel von PDF/A

ISO 19005-1 definiert "ein Dateiformat basierend auf PDF, genannt PDF/A, welches einen Mechanismus zur Verfügung stellt, um elektronische Dokumente auf eine Weise darzustellen, so dass das visuelle Erscheinungsbild über die Zeit erhalten bleibt, unabhängig von den Werkzeugen und Systemen zur Herstellung, Speicherung und Reproduktion." (Auszug aus ISO 19005-1). Der Standard definiert nicht eine Archivierungs-Strategie und gibt auch nicht die Ziele der Archivierung vor. Er beschreibt höchstens die Anforderungen an die Form von elektronischen Dokumenten, um sicherzustellen, dass sie in den nächsten Jahren einwandfrei reproduziert werden können.

Ein Schlüsselfaktor für die Reproduzierbarkeit von PDF/A Dokumenten ist, dass alle notwendigen Informationen im Dokument selbst enthalten sind. Dies umfasst sichtbaren Inhalt wie Texte, Vektorgraphiken, Rasterbilder, Schriftarten, Farbräume und vieles mehr. Hingegen darf ein PDF/A Dokument weder direkt noch indirekt auf externe Quellen verweisen. Beispiele davon sind Verbindungen zu vorgehaltenen Bildern oder nicht im PDF/A-Dokument selbst eingebettete Schriften.

PDF/A-Dateien sind selbstbeschreibend.

Alle benötigten Informationen zur Anzeige eines Dokumentes sind in der Datei eingebettet.

Zur Zeit basiert PDF/A auf der PDF-Referenz Version 1.4.

Ein Vergleich von PDF mit PDF/A

PDF an sich garantiert keine Langzeit-Reproduzierbarkeit, nicht einmal das Prinzip WYSIWYG (*what you see is what you get*). Damit beides gewährleistet ist, mussten gewisse Einschränkungen und Erweiterungen in den Standard aufgenommen werden. Ferner, um bei einem breiten Publikum akzeptiert zu werden, musste PDF/A auf einer bereits existierenden PDF-Version aufbauen. Das ISO TC 171 hat Adobe's PDF Referenz 1.4 (von Adobe implementiert in der Acrobat 5-Version) als Grundlage des Standards gewählt. Der ISO-Standard sagt aus, dass PDF/A „alle Anforderungen der PDF Referenz erfüllen muss, wie durch diesen Teil des ISO 19005 Standards ergänzt". Der Standard beschreibt also nur die Unterschiede zur Referenz. Um PDF/A vollständig zu verstehen, muss also auch die PDF Referenz 1.4 verstanden werden.

Bestimmte, in PDF 1.4 erlaubte Funktionalität wie die Transparenz oder die Ton- und Videoreproduktion, sind aus PDF/A ausgeschlossen worden. Es gibt andererseits in PDF 1.4 optionale Konstrukte, welche in PDF/A vorhanden sein müssen. So müssen in PDF/A beispielsweise alle verwendeten Schriften eingebettet sein. Kurzum, PDF/A präzisiert im Wesentlichen spezifische Eigenschaften der PDF Referenz 1.4 und definiert ob sie obligatorisch, empfohlen, eingeschränkt oder verboten sind.

Das PDF/A, A-1a, A-1b, A-2 "Babylon"

PDF/A ist als mehrteilige Standardreihe angelegt. Bislang ist nur PDF/A-1 (Part 1) verabschiedet worden. PDF/A-1 wiederum ist weiter unterteilt in die Übereinstimmungsgraden PDF/A-1a und PDF/A-1b.

PDF/A-1a (Level A Conformance) bezeichnet die vollständige Übereinstimmung mit dem PDF/A Standard (ISO 19005-1): Part 1.

Es existieren verschiedene Stufen von PDF/A.

PDF/A-1a bedeutet die vollständige Übereinstimmung mit der PDF/A Norm.

Die Beschreibung der Mindestanforderungen zur Übereinstimmung mit PDF/A ist in PDF/A-1b (Level B Conformance) enthalten. Die PDF/A-1b Anforderungen sollten für die visuelle Langzeit-Reproduktion genügen.

PDF/A-1a und PDF/A-1b unterscheiden sich hauptsächlich bezüglich Durchsuchbarkeit (Textextraktion).

- PDF/A-1a stellt sicher, dass die logische Dokumentenstruktur und die natürliche Leseordnung des enthaltenen Texts erhalten bleibt. Die Textextraktion ist vor allem dann wichtig, wenn die Dokumente beispielsweise auf mobilen Geräten (wie beispielsweise einem PDA) oder auf Geräten gemäss Abschnitt 508 des „US Rehabilitation Act" dargestellt werden sollen. In solchen Fällen muss der Text auf dem eingeschränkten Bildschirm neu angeordnet werden können (re-flow). Diese Eigenschaft ist auch unter der Bezeichnung „Tagged PDF" bekannt.
- PDF/A-1b stellt sicher, dass der Text (wie auch der übrige Seiteninhalt) korrekt angezeigt werden kann, garantiert aber nicht, dass der enthaltene Text auch lesbar und verständlich ist. Damit ist die Übereinstimmung mit dem oben genannten Abschnitt 508 nicht gewährleistet.

Der Unterschied zwischen PDF/A-1a und PDF/A-1b hat in der Regel keine Relevanz für gescannte Dokumente, sofern sie nicht mittels OCR für die Suche angereichert wurden.

Zur Zeit arbeitet das technische Komitee an einem neuen Teil des Standards: ISO 19005-2, Part-2 (PDF/A-2). PDF/A-2 soll auf die Eigenschaften der neueren Versionen der PDF Referenz 1.5, 1.6 und 1.7 eingehen. PDF/A-2 soll aufwärts-kompatibel sein, d.h. alle gültigen PDF/A-1-Dokumente sollen gleichzeitig mit PDF/A-2 übereinstimmen. Hingegen entsprechen PDF/A-2 Dokumente nicht mehr notwendigerweise dem PDF/A-1 Standard.

PDF/A- 1b gewährleistet nicht die Textextrahierung (wie im Abschnitt 508 des "US Rehabilitation Act" verlangt).

PDF/A-2 wird zur Zeit erarbeitet und berücksichtig die Versionen 1.5, 1.6 und 1.7 der PDF-Referenz.

Der PDF/A-Standard

Wie erhalte ich eine Kopie?

Der ISO 19005-1 Standard kann auf der ISO Web Seite gekauft werden. Kopien können als Papier oder elektronisch im PDF-Format bestellt werden und sind, wie alle anderen ISO Standards, mittels Copyright geschützt. Deshalb ist die Bereitstellung frei erhältlicher Kopien über das Internet nicht legal. Der Standard ist zur Zeit nur in englischer Sprache verfügbar.

Wer sollte den Standard lesen?

Der PDF/A Standard beabsichtigt die Unterstützung und Verbesserung von Archivierungs-Strategien. Der Standard selbst ist sehr technisch gehalten und kann nur von Experten mit fundiertem Wissen über Seitenbeschreibungssprachen wie PostScript und PDF vollständig verstanden werden. Der Umfang des Hauptdokuments ist zwar klein, der Umfang der Grundlagen-Dokumente jedoch sehr gross. Allein die PDF Referenz 1.4 umfasst fast 1000 Seiten, die darin referenzierten Dokumente wie Fontformate, XML-Spezifikation, Kompressionsformate, RFCs usw. nicht eingeschlossen. Zudem garantiert der Standard alleine nicht die Langzeit-Archivierung. Man ist gut beraten einen Experten hinzuzuziehen, um die PDF/A-Anforderungen zu verstehen, daraus eine unternehmensweite Archivierungs-Strategie herzuleiten und die langfristigen Ziele für die Aufbewahrung von Dokumenten zu erreichen.

Die PDF/A-Norm (ISO 19005-1) kann erworben werden auf: www.iso.org.

PDF/A ist komplex. Allein die PDF Referenz 1.4 beinhaltet knapp 1'000 Seiten.

Welche Werkzeuge gibt es?

Werkzeuge für die Erzeugung, Verarbeitung und Validierung von PDF/A-Dokumenten sind seit Mitte 2006 auf dem Markt. Von Adobe selbst wurden entsprechende Funktionen in die im Herbst 2006 erschienen Version 8 von Adobe Acrobat integriert. Auch Microsoft stellt für das neue Office 2007-Paket ein separat herunterzuladendes Plug-In bereit, um PDF/A-konforme Dateien direkt aus Office-Produkten heraus erzeugen zu können. Angesichts der bereits erschienenen Produkte zum Erstellen von PDF/A ist es mittlerweile sehr wichtig geworden, die jeweils erstellten PDF/A-Dokumente bezüglich der einwandfreien PDF/A-Konformität zu überprüfen.

PDF/A verlangt nach einer vollständigen Lösung

PDF/A ist nur ein Teil einer vollständigen Archivierungslösung. PDF/A alleine garantiert keine Langzeit-Archivierung und garantiert nicht, dass die Anzeige so funktioniert, wie beabsichtigt. PDF/A beansprucht für sich auch nicht, in jedem Fall die adäquateste Lösung zu sein. Hingegen definiert PDF/A spezifische Anforderungen an elektronische Dokumente, um die Langzeit-Archivierung zu ermöglichen. Wenn ein PDF/A-konformes Archiv implementiert werden soll, müssen noch weitere Aspekte berücksichtigt werden. Dazu gehören beispielsweise unternehmensweite Standards und Prozeduren, verlässliche Datenquellen, Qualitätsmanagement und spezifische, auf den Anwendungszweck zugeschnittene Anforderungen. Vor allem die Migration der bestehenden Papier- oder TIFF-Archive in PDF/A-konforme Archive ist keine unbedeutende Aufgabe und muss deshalb sorgfältig geplant werden.

PDF/A ist nur ein Element einer kompletten Archivierungsstrategie.

PDF/A allein gewährleistet nicht die Langzeitarchivierung, aber es ermöglicht sie.

Zusammenfassung

PDF/A als neuer Archivierungs-Standard

Die Erwartung an PDF/A ist, sich als neuer Standard für die Archivierung von elektronischen Dokumenten zu etablieren. PDF ist weltweit omni-präsent im öffentlichen und privaten Sektor und bereits als Format für unzählige Zwecke akzeptiert. Der PDF/A-Standard hilft nun sicherzustellen, dass Benutzer diese Dokumente auch nach langer Zeit noch sicher reproduzieren können.

Die Einführung des PDF/A-Standards wird vermutlich (und sollte es auch) einen Einfluss auf die zukünftige Entwicklung von PDF selbst haben. Adobe wird unabhängig davon mit der Verbesserung und Einführung neuer Eigenschaften fortfahren. Beispiele dafür sind 3D-Modelle oder XFA für dynamische PDF-Formulare. Dies wird weiteren Druck auf den „Standard" ausüben, weil das Wesen eines Standards - insbesondere eines Archivierungs-Standards - darin besteht, dass man ihn nicht häufig ändert.

Wie wird der Markt reagieren?

Wir sollten nicht erwarten, dass der Markt nun geradezu mit PDF/A-Produkten überflutet wird. Es braucht beträchtliches Wissen, um die Technologie hinter PDF/A zu verstehen. Zudem hat der Anwender höhere Qualitätsanforderungen an Standard-konforme Software. Erste Werkzeuge sind seit Mitte 2006 auf dem Markt erschienen. Gefragt sind die PDF/A-konforme Erzeugung, die PDF/A-Validierung sowie eine einfache Konversion von bestehendem PDF in Standard-konformes PDF/A.

Das Erscheinen der ersten, professionellen PDF/A Werkzeuge löst bereits auch Prozesse zur Implementierung von PDF/A-konformen Archivsystemen aus. Man darf gegenwärtig auch noch nicht zu viel Funktionalität erwarten. Es ist wahrscheinlich, dass vorerst nur das eingeschränkte PDF/A-1b angeboten wird und das vollumfängliche PDF/A-1a erst später. Wie oft zu beobachten, wird es auch eine Reihe von Produkten geben, welche PDF/A-Konformität für sich beanspruchen, es in Wirklichkeit aber nicht sind. Expertise für die Evaluation und seriöse Anbieter sind vor allem in der Markteinführungsphase gefragt.

Heisse Luft oder eine langfristige Strategie?

PDF/A ist bestimmt nicht eine Eintagsfliege. Das Bedürfnis nach einem Standard für die Archivierung auf der Grundlage von PDF ist schon einige Jahre alt. Und: PDF wird bereits, mit Hilfe unternehmensspezifischer Richtlinien, in vielen Anwendungen für diesen Zweck verwendet. Dass Microsoft ihren Kunden nachgegeben hat und die direkte Erzeugung von PDF/A aus ihren neuesten Office-Produkten heraus unterstützt, ist schon für sich genommen ein klares Signal. PDF/A, international akzeptiert, wird von Dauer sein.