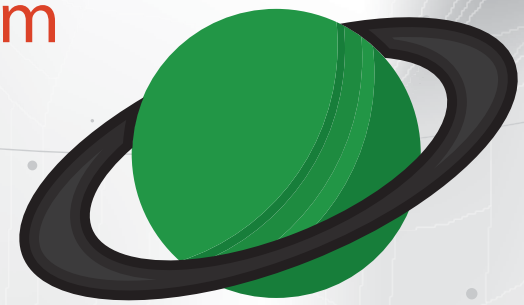
Juli / August 2013 Ausgabe #3-13 Deutschland Euro 12,00 ISSN: 1864-8398 www.dokmagazin.de

DOK.

Technologien, Strategien & Services für das digitale Dokument

Potenziale von SharePoint 2013

als Collaboration-Plattform



Der Kunde im Fokus: Social Media Management

Special **Archivieren mit Methode**
Von Microfilm- bis Web-Archivierung

Requirements Engineering für Druckdokumente

Im Trend: Versionsmanagement ist angesagt!

Digital born PDF/A – Knacknuss oder (v)erkanntes Potenzial?

PDF/A-Archivierung, Dokumentenformate, Konvertierung, Validierung

Die Archivierung gescannter Dokumente in PDF/A wird seit mehr als sechs Jahren erfolgreich praktiziert. Mit der Archivierung digital erzeugter Dokumente ist man nach wie vor zurückhaltend. Was sind die Gründe dafür? Einige sind offensichtlich: Gescannte Dokumente sind einfacher in PDF/A zu konvertieren, die Umwandlung digital erzeugter Dokumente ist hingegen meist eine technische Herausforderung. Etwas weniger offensichtlich sind Fehler in der Wiedergabe des konvertierten Dokuments, Funktionseinschränkungen des PDF/A-Standards und weitere Gründe. Mit den richtigen Strategien lassen sich die Herausforderungen jedoch meistern.

Noch entsteht ein großer Teil des elektronischen Archivguts aus gescannten Dokumenten wie Geschäftskorrespondenz, Belege für die Buchhaltung, Verträge, aus Papierarchiven und andere aufbewahrungswürdige Papiere, welche in ihre elektronische Form migriert werden sollen. Die Anzahl der elektronisch erzeugten Dokumente holt aber rasch auf, meist Rechnungen aus ERP-Systemen, E-Mails, Office-Dokumente im Postausgang aber auch speziellere Dokumente wie Konstruktionszeichnungen aus CAD-Systemen.

Abbildungstreue – technische Herausforderung bei gescannten Dokumenten

Es ist eine Tatsache: Gescannte Dokumente sind im Wesentlichen Rasterbilder. Jahrelang war es in Ordnung, diese als TIFF aufzubewahren, meist in Schwarz und Weiß, um Speicherplatz zu sparen. Die Anforderungen sind jedoch gestiegen. Farbe, Metadaten und Volltextsuche sind heute durch den ISO-Standard PDF/A selbstverständlich, ohne wesentlich mehr ▶

www.pdf-tools.com

Dr. Hans Bärfuss ist Gründer und Geschäftsführer der **PDF Tools AG** und Delegierter der Schweizerischen Normenvereinigung (SNV) bei der ISO. Er ist auch einer der Initianten und Gründer der PDF Association und heute Chairman des Swiss Chapter. Die PDF Tools AG ist ein Hersteller von Softwarelösungen und Programmierkomponenten für die PDF und PDF/A Erzeugung, Bearbeitung, Wiedergabe und Archivierung.

Speicherplatz zu verbrauchen. Die technischen Herausforderungen im Zusammenhang mit diesen Rasterbildern konzentrieren sich auf die Bildanalyse und Verarbeitung. Dazu gehören:

- Die Bilder werden durch eine Texterkennungsmaschine (OCR) weiterverarbeitet. Leere Seiten werden erkannt, das Bild wird entfleckt und gerade gerichtet. Danach erfolgt die Erkennung des Textes und des Barcodes.
- Segmentierung und Kompression: Das Farbbild jeder Seite wird in seine Bestandteile wie Hintergrund, Text und Fotos zerlegt. Die einzelnen Teile werden durch spezifisch dafür entworfene Kompressionsverfahren in der Größe reduziert. Dieses Mixed Raster Content-Verfahren (MRC) ermöglicht Farbdokumenten, konkurrenzfähige Dateigrößen zu Schwarz und Weiß zu erreichen.

Diese Verfahren haben die Softwarehersteller schon vor der PDF/A-Area zu beherrschen gelernt. Mit PDF/A erhält man jedoch – im Unterscheid zu TIFF – ein standardisiertes Resultat. PDF/A, als Untermenge von PDF, kann jedoch viel mehr. Mit seinen Farbräumen, Schriften, Vektoren, Füllmustern und Transparenzmischungen verfügt PDF über eines der mächtigsten 2D-Grafikmodelle und ist geradezu prädestiniert, um digital erzeugte Dokumente wiederzugeben. Man muss nur noch die digitale Quelle in PDF/A konvertieren. Allerdings ist dieser Schritt eine größere technische Herausforderung als es auf den ersten Blick scheint.

Zunächst gibt es die große Anzahl von Dokumentenformaten, welche umgewandelt werden sollen: ASCII-Texte, Word, Excel, PowerPoint, PDF, E-Mails, HTML und XML von verschiedenen Orten wie Dateiablagen, ZIP-Archiven, Mailboxen, Dateianhänge und Datenströme aus Applikationen. Zudem reicht die Qualität der digitalen Quelle meist nicht an die von Rasterbildern heran. Die Dateien sind entweder auf dem Übermittlungsweg beschädigt oder von Anfang an schlecht erzeugt worden. Gerade bei PDF-Dateien, welche mit Freeware erzeugt wurden, ist dies sehr oft der Fall. Das Problem der „Bad-PDF" verursacht nicht nur bei Softwareherstellern hohe Kosten, sondern führt auch immer öfter zu Problemen in dokumentenbasierten Geschäftsprozessen.

Die größte Herausforderung bei der Umwandlung von Dokumenten aus digitalen Quellen in PDF/A ist jedoch die Abbildungstreue. Auch wenn die konvertierte Datei formal einwandfrei dem ISO-Standard genügt, kann es vorkommen, dass das visuelle Resultat nicht dem Original entspricht. Solche Abbildungsfehler können viele Ursachen haben. Meist liegt es daran, dass die Quelldokumente eine hohe grafische Komplexität wie beispielsweise Füllmuster oder Transparenz aufweisen und die Umwandlungssoftware nicht alle Grafikfunktionen oder alle Kombinationen davon in PDF/A abbilden kann. Ein typisches Beispiel sind die vielen virtuellen Druckertreiber, welche zur Erzeugung von PDF/A-Dateien über die Druckfunktion dienen. Die meisten dieser Treiber stützen auf den vom Betriebssystem mitgelieferten PostScript-Treiber ab, welcher jedoch nur einen Teil der definierten Grafikschnittstelle implementiert.

Strategien für fehlerfreie PDF/A-Dokumente

Heute ist es keine Grundsatzfrage mehr: PDF/A ist als Archivformat sowohl für gescannte als auch für digital erzeugte Dokumente geeignet. Allerdings übt man sich bislang aufgrund der genannten technischen Schwierigkeiten bei der Umwandlung digitaler Quellen in PDF/A in vorsichtiger Zurückhaltung. Doch lassen sich diese Herausforderungen meistern. Dabei spielt die Wahl der Konversionssoftware eine wichtige Rolle – die Wahl der richtigen Systemarchitektur ist jedoch darüber hinaus entscheidend für den Erfolg.

Bei gescannten Dokumenten hat es sich bewährt, dass die Umwandlung des gescannten Bildes in ein durchsuchbares, mit Metadaten angereichertes und möglicherweise digital signiertes Dokument in einer dafür spezialisierten Software (Scan Server) geschieht. Dabei sind alle Verarbeitungsschritte optimal aufeinander abgestimmt. Es ist wichtig, dass der Scanner nur das rohe Bild liefert, um eine optimale Kompressionsleistung zu ermöglichen. Wird die Verarbeitung auf den Scanner, den Scan-PC und den Server verteilt, ist das Ergebnis meist suboptimal.

Für eine professionelle Umwandlung digital erzeugter Dokumente in PDF/A gibt es verschiedene Wege. Der einfachste Fall ist, wenn das Dokument bereits in PDF/A erstellt wurde – wie beispielsweise Angebote, Rechnungen oder Berichte. Dann muss nur noch mit Hilfe eines Werkzeugs (PDF/A-Validator) geprüft werden, ob das Dokument die Regeln des Standards einhält.

Ist das Quelldokument kein PDF/A, muss es umgewandelt werden. Im besten Fall bietet dann die native Applikation, wie beispielsweise ein Produkt der Microsoft Office-Palette, eine direkte Funktion („Save as PDF/A“) an. Die Erfahrung hat allerdings gezeigt, dass diese Funktionen Abbildungsfehler und kleinere Verstöße gegen den PDF/A-Standard aufweisen. Eine bewährte Strategie ist daher, die weniger heikle Funktion für die direkte Erzeugung einer gewöhnlichen PDF-Datei („Save as PDF“) zu verwenden. Das Resultat wird anschließend mit einem spezialisierten Konverter in PDF/A umgewandelt.

Steht eine direkte Funktion zur PDF/A-Erzeugung nicht zur Verfügung, bleibt oft nur den Weg über die Druckfunktion. Das Dokument wird über einen virtuellen Druckertreiber auf eine PDF/A-Datei „ausgedruckt“. Hier ist zu empfehlen, dass ein speziell dafür entwickelter PDF/A-Druckertreiber eingesetzt wird,

um Abbildungsfehler zu vermeiden, wie sie bei üblichen PDF-Druckertreibern auftreten, welche auf PostScript basieren.

Zentrale PDF/A-Konversionslösung – der sichere Weg

Um es gleich vorweg zu nehmen: Eine zentrale PDF/A-Konversionslösung, sowohl für gescannte als auch für digital erzeugte Dokumente, lohnt sich schon für wenige Arbeitsplätze. Die Gründe dafür sind einfach:

- Qualität: Durch die geschützte Laufzeitumgebung auf dem Server kann sichergestellt werden, dass alle Verarbeitungsschritte im Umwandlungsprozess immer gleich und mit den optimal dafür ausgewählten Werkzeugen ausgeführt werden.
- Unterstützte Formate: Zentrale Lösungen können eine Vielzahl von Dokumentenformaten unterstützen, auch für Formate, für die auf dem Client keine geeignete Software existiert. Damit erspart man sich das aufwändige Ausrollen von Software auf die Arbeitsstationen.
- Robustheit und Stabilität: Die Applikationen zur Umwandlung werden in einer automatisierten und überwachten Laufzeitumgebung betrieben. Dadurch kann sichergestellt werden, dass der Konversionsdienst immer zuverlässig zur Verfügung steht. Der Server überwacht das korrekte Funktionieren der Applikationen und startet sie automatisch bei auftretenden Problemen neu.
- Validierung: Die erzeugten Dateien werden vom Server auf Konformität mit dem Standard geprüft. Zusätzlich kann der Server zur Sicherheit einen automatischen Bildvergleich vornehmen, um Abbildungsfehler auszuschließen.
- Skalierbarkeit: Konversionsserver können einerseits durch Multiprozessor-Maschinen oder die Verteilung auf mehrere Maschinen einfach skaliert werden.
- Zentralisierung: Da der Server zentral betreut und die Clients schlank ausstattet werden können, lassen sich Betriebskosten sparen.

Alles in allem also überzeugende Argumente, die Konversion in PDF/A mithilfe professioneller Tools durchzuführen.

Fazit

Der PDF/A-Standard wird laufend weiterentwickelt und neuen Bedürfnissen angepasst. Die Umwandlung gescannter und digital erzeugter Dokumente in PDF/A ermöglicht vielen Unternehmen in unterschiedlichsten Branchen, den wachsenden Anforderungen an eine sichere digitale Archivierung gerecht zu werden und auch langfristig jederzeit auf die Dokumente zugreifen zu können. Durch die Anwendung bewährter Strategien steht der Umsetzung eines erfolgreichen digitalen Archivierungsprojektes, welches die technischen, rechtlichen und betriebswirtschaftlichen Aspekte erfüllt, nichts im Wege.