Formatierend. Praktisch. Gut.

Digitale Langzeitarchivierung, PDF/A-Erzeugung, Scan Server, Texterkennung, Metadaten, Digitale Signatur, Datenkompression

www.pdf-tools.com

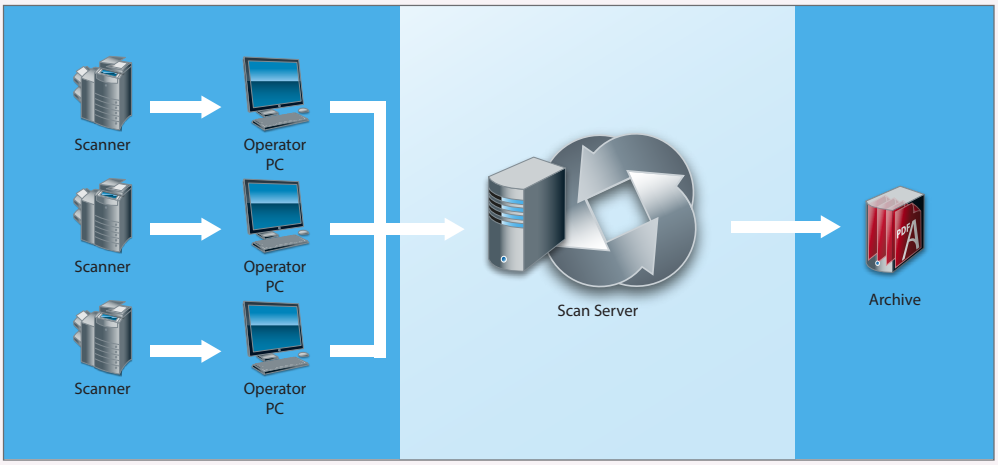
Dr. Hans Bärfuss ist Gründer und Geschäftsführer der **PDF Tools AG**, einer international erfolgreichen Softwareentwicklungs- und Vertriebsgesellschaft. Er ist Delegierter der Schweizerischen Normenvereinigung (SNV) bei der ISO und hilft bei der Standardisierung von Dateiformaten und digitalen Signaturen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

Fensterlose Räume mit Papierakten zu füllen und Mitarbeiter mit dem Suchen von Papierdokumenten zu beschäftigen sind zu Zeit- und Kostenfaktoren geworden, die sich kaum jemand mehr leisten möchte. Ein digitales Archiv muss her, heißt es in den Führungsetagen nicht nur großer Unternehmungen. Aber wie? Überlasst das den Herstellern der Scan-Geräte, sagen die einen. Dazu braucht es mehr als das, sagen die anderen.

Braucht man mehr als einen Scanner?

Das Scannen von Papierdokumenten im Posteingangsbereich einer Unternehmung ist zum Alltag geworden. Je nach Art und Menge der anfallenden Papierdokumente werden dafür Multifunktionsgeräte (MFP) oder Hochleistungsscanner eingesetzt. In den meisten Fällen werden die gescannten Bilder als TIFF-Dateien in Schwarz und Weiß erzeugt, so wie man dies von den FAX-Maschinen gewohnt ist. In speziellen Anwendungen wie dem Scannen von Schecks, Fotos für Ausweise usw. wird die Datei in Farbe erzeugt. Allerdings verzichtet man oft auf das Scannen in Farbe, weil die resultierenden TIFF-Dateien entweder zu groß sind oder die verwendete JPEG-Kompression die Bildqualität sichtbar reduziert.

Eine gute Bildqualität ist jedoch eine wichtige Voraussetzung für eine gute Texterkennungsrate. Für eine hohe Kompressionsrate bei gleichzeitig guter Bildqualität benötigt man Rechenleistung, die in den dezentralen Multifunktionsgeräten selbst oft nicht zur Verfügung steht. Für diesen Aspekt bietet eine separate Scan-Software entscheidende Vorteile.



Scan Server – zentraler Dienst für die Erzeugung von PDF/A-Dateien aus gescannten Dokumenten

Die einzelnen Bearbeitungsschritte wie Texterkennung, Kompression, PDF/A-Erzeugung und Digitale Signatur können in der Regel nicht durch den Scanner alleine ausgeführt werden, weil oft nachträglich Metadaten von einer Index-Station hinzugefügt werden. Dieser Arbeitsschritt bricht jedoch das Siegel der digitalen Signatur und macht sie wertlos. Auch für diesen Aspekt bietet eine separate Software einen entscheidenden Vorteil.

PDF/A – ein universeller Dokumentenstandard?

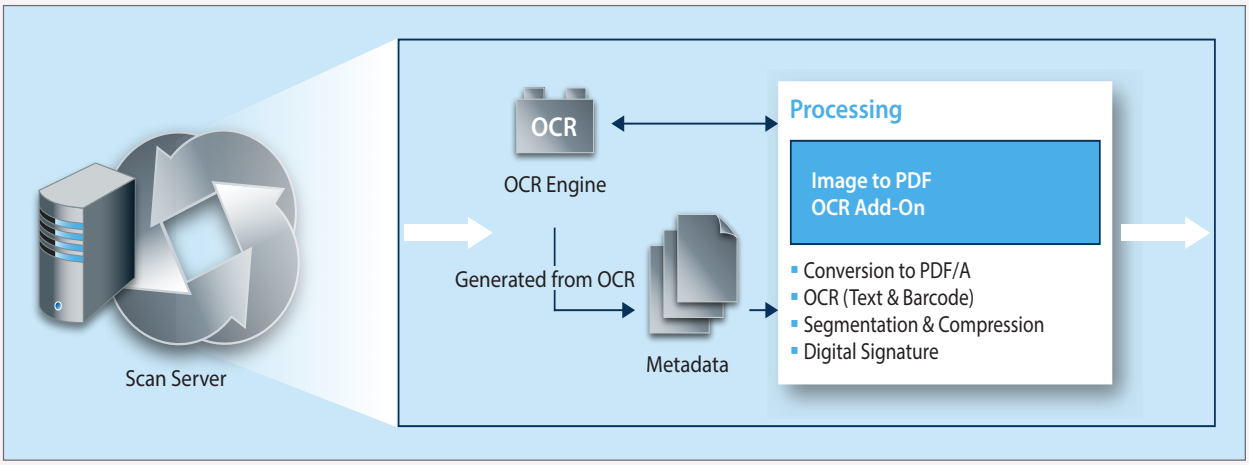
Der PDF/A-Standard hat sich heute in Posteingangs-Anwendungen weitgehend durchgesetzt. Die wichtigsten Vorteile des PDF/A-Standards gegenüber den klassischen Dokumentenformaten wie TIFF und JPEG sind:

- Einheitliches Format: PDF/A ist für die Speicherung sowohl von gescannten als auch von digital erzeugten Dokumenten gleichermaßen geeignet.
- Hohe Kompressionsrate: Der PDF/A-Standard unterstützt modernere und leistungsfähige Kompressionsverfahren und somit auch kleine Dateigrößen für Farbbilder.
- Texterkennung: Die erzeugten PDF/A-Dokumente können durch das Einbetten von Texten aus einer OCR-Maschine durchsuchbar gemacht werden.
- Eingebettete Metadaten: Damit das Dokument und die dazugehörenden Metadaten eine unteilbare Einheit bilden, werden in PDF/A die Metadaten in die Datei eingebettet. PDF/A verwendet für die Speicherung das Extensible Metadata Platform (XMP) Format, welches unabhängig von PDF/A als eigener ISO Standard definiert ist.
- Digitale Signatur: Um die Integrität und Authentizität der erzeugten Dokumente zu gewährleisten, kann optional eine digitale Signatur nach dem PAdES-Standard auf das PDF/A-Dokument aufgebracht werden. Die digitale Signatur ist eine Form der elektronischen Signatur, welche dem Erfordernis der handschriftlichen Unterschrift gleich gerecht werden kann, wie die handschriftliche Unterschrift selbst, sofern die gesetzlichen Voraussetzungen (nationale Signaturgesetze) dafür erfüllt sind.

Anforderung	TIFF	PDF/A
Langfristige Lesbarkeit	+	+
Eindeutige Wiedergabe	+	+
Datenkonsistenz	Proprietäre Tags für Metadaten	+
Authentizität / Integrität	Mit abgesetzten Signaturen	+
Speicherplatzbedarf	Schwarz / Weiss: + Farbe: –	+
Durchsuchbarkeit	Proprietäre Tags für OCR Text	+
Langzeiterfahrung	+	+

Die Vorteile von PDF/A gegenüber TIFF

Alle diese Vorteile lassen sich mit TIFF-Dokumenten grundsätzlich auch realisieren, jedoch nur als proprietäre Erweiterungen, da der TIFF-Standard selbst dafür keine Lösungen bereithält.



Hauptfunktionen des Scan-Servers

Was kann ein zentraler Scan-Server?

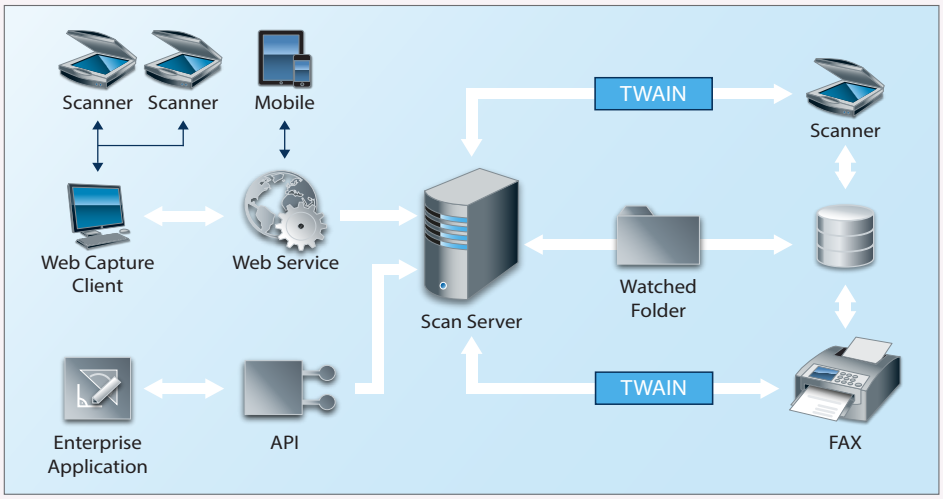
Ein Scan-Server ist ein zentraler Dienst, welcher in einer Unternehmung dezentral gescannte Dateien und dazugehörende Indexdaten in das standardisierte Dateiformat PDF/A umwandelt. Dafür übernimmt der Dienst alle Aufgaben, welche von den dezentralen Scan-Stationen an ihn delegiert werden können. Besonders geeignet sind Verarbeitungsschritte, welche keine Benutzerinteraktion erfordern oder die Effizienz der lokalen Scan-Stationen mit leistungsintensiven Funktionen (OCR, Kompression) belasten.

Die Hauptfunktionen des Dienstes sind:

- Text- und Barcodeerkennung: Gescannte Bilddateien sollen durchsuchbar gemacht werden. Der Dienst kann einen Texterkennungsdienst nutzen, um den Text in der Bilddatei zu erkennen und in die umgewandelte Datei so einzubetten, dass diese durchsuchbar wird. Die erkannten Barcodes können mehrfach verwendet werden, in der Textsuche, als Teil der eingebetteten Metadaten und zur Steuerung der Verarbeitung (Name der Ausgabedatei, Seitentrennung usw.) im Dienst.
- Kompression: Farbbilder werden in mehrere Ebenen zerlegt und durch das Mixed-Raster-Content -Verfahren (MRC-Verfahren) stark und ohne sichtbare Verluste komprimiert.
- Einbettung von Metadaten: Der PDF/A-Standard sieht vor, dass Metadaten in Form von XMP-Paketen in das Dokument eingebettet werden. Der Dienst bietet diese Funktion an.
- Erzeugen einer PDF/A-Datei: Der Dienst erzeugt ein- oder mehrseitige Ausgabedokumente entsprechend der ISO-19'005 Standardreihe. Alle zurzeit veröffentlichten Normteile PDF/A-1, PDF/A-2 und PDF/A-3 werden unterstützt.
- Digitale Signatur: Die Signatur kann fortgeschritten oder qualifiziert sein, für die Langzeitaufbewahrung oder nur für den Austausch geeignet sein und wahlweise einen Zeitstempel enthalten. Anstelle der persönlichen Signatur kann auch nur ein Zeitstempel aufgebracht werden. Der Dienst kann eine kryptografische Infrastruktur (USB Token, HSM) über eine Standardschnittstelle (PKCS#11) nutzen, um digitale Signaturen zu erzeugen.

Ein typischer Ablauf sieht wie folgt aus:

- Bildakquisition: Der Scan-Operator startet den Scan-Vorgang und erzeugt eine TIFF-Datei in Farbe. Der Scanner legt Dateien in der Regel in einem Dateiordner ab. Faksimile-Dokumente werden von der FAX-Maschine empfangen und als TIFF-Dateien in Schwarz und Weiß in einen speziellen Ordner abgelegt.
- Manuelle Klassifikation: Wahlweise und je nach Prozess führt der Scan-Operator eine Klassifikation durch. Er steuert den Scanner dabei so, dass die Bilder in verschiedenen Dateiordnern abgelegt werden (z. B. für Rechnungen oder Lieferscheine) oder spezielle Barcode-Blätter einfügt, welche für die Trennung und Klassifikation der Dokumente dienen oder er erfasst einen minimalen Satz von Indexdaten.
- Segmentierung und Kompression: Das Farbbild jeder Seite wird in seine Bestandteile wie Hintergrund, Text und Bilder zerlegt. Die einzelnen Teile werden durch spezifisch dafür entworfene Kompressionsverfahren in der Größe reduziert. Dieses MRC-Verfahren ermöglicht Farbdokumenten, konkurrenzfähige Dateigrößen zu erreichen.



Anwendungen des Scan-Servers

- Text- und Barcode-Erkennung: Die Bilder werden durch eine OCR-Maschine weiterverarbeitet. Als Erstes wird das Bild entfleckt und geradegerichtet, danach erfolgt die Erkennung der Texte und der Barcodes.
- Metadaten: Informationen aus der manuellen Klassifizierung, der erkannten Barcodes und weiteren Quellen werden zu standardisierten XMP-Metadaten zusammengefügt.
- PDF/A-Erzeugung: Die aufbereiteten Bilder jeder Seite, der erkannte Text und die Metadaten werden zusammen mit dem ICC-Farbprofil des Scanners zu einem PDF/A-Dokument zusammengefügt. Optional kann eine Index-Datei erzeugt werden, welche nur die Metadaten enthält.
- Digitale Signatur: Wahlweise kann eine digitale Signatur aufgebracht werden, damit die Nachvollziehbarkeit und Revisionsfestigkeit des Dokuments sichergestellt ist.
- Validierung: Wahlweise können die PDF/A-Konformität des erstellten Dokumentes und die Gültigkeit der digitalen Signatur überprüft werden.

Ein solcher Dienst bietet in der Regel auch eine Reihe von Zusatzfunktionen an.

Wo wird der Dienst eingesetzt?

Ein Scan-Server wird für die folgenden Anwendungen eingesetzt:

- Paper Capture: Elektronische Archivierung von Papierdokumenten, welche im Posteingang eines Unternehmens anfallen.
- Facsimile Capture: Elektronische Archivierung des gesamten FAX-Verkehrs zwischen dem Unternehmen und seinen Geschäftspartnern.
- Archive Migration: Migration von Papierarchiven in ein elektronisches Archiv mit dem standardisierten PDF/A-Format.
- Web/Mobile Capture: Nutzung des zentralen Dienstes in Client-/Server-Anwendungen über einen Webdienst
- Enterprise Application Integration: Nutzung des zentralen Dienstes für die PDF/A-Dokumentenerzeugung über eine Programmierschnittstelle (API) aus Fachapplikationen heraus, welche TIFF- oder JPEG-Dateien erzeugen.

Fazit

Der Aufbau eines digitalen Langzeitarchivs ist für große Unternehmen zu einem Muss geworden. Es lohnt sich aber auch bereits für mittlere und kleinere Unternehmen, um Lager- und Personalkosten zu sparen.

Ein gut durchdachter Scanning-Prozess hilft so früh wie möglich, nämlich bereits beim Posteingang, das lästige Papier loszuwerden. Digitale Unterschriften sorgen dabei für den Erhalt der Beweiskraft der elektronischen Dokumente. Ein zentraler Scan-Dienst hilft dabei, einen leistungsfähigen, flexiblen und zukunftssicheren Verarbeitungsprozess zu implementieren.

PDF/A, als standardisiertes Dateiformat für das Langzeitarchiv, ist nicht nur für gescannte Dokumente geeignet, sondern dient gleichermaßen als universelles Format für digital erzeugte Dokumente.