PDF Tools: Digitalisierung von Dokumenten in PDF/A

Wie komplexe Projekte gelingen

Öffentliche und private Unternehmen geben sich zeitgemäß, lancieren Projekte und reiten auf der Digitalisierungswelle. Infrastrukturen für die Zentralisierung von Archiven und die weltweite Online-Recherche werden geschaffen, damit, wie immer bei solchen Vorhaben, die Produktivität gesteigert und die Kosten gesenkt werden. Aber haben wir denn die Digitalisierung im Griff? Schaffen wir damit nicht neue Risiken? Was muss man darüber wissen, damit die Projekte nicht zum Albtraum werden? Ein Beitrag von Dr. Hans Bärfuss, Geschäftsführer der PDF Tools AG.

Dr. Hans Bärfuss, Gründer und Geschäftsführer der PDF Tools AG: „Die Frage ist, wie die unterschiedlichen Bedürfnisse in einem Unternehmen zu einer einheitlichen Scan-Strategie vereint werden sollen."

Damit Digitalisierungsprojekte nicht zum Albtraum werden, braucht es geeignete Scanner und passende Software. Vielleicht auch einen guten Berater, um einen möglichst idealen Workflow zu definieren. Für ein gutes Gelingen der Projekte ist jedoch grundlegendes Wissen über den Prozess der Digitalisierung bestimmt hilfreich. Dieser Artikel gibt einen Einblick in einige Aspekte von spezialisierter Software für die Digitalisierung und soll damit deren Auswahl und Einsatz in konkreten Projekten erleichtern.

Was macht die Scan-Software?

Auf dem Weg vom Papier zum archivfähigen Dokument führt die Scan-Software, unabhängig von Architektur und Umfang, eine Reihe von Verarbeitungsschritten aus, auch wenn einzelne dieser Stufen wahlweise ausgeführt werden oder für den Benutzer nicht sichtbar sind.

Bildakquisition: Der Scanner erzeugt aus dem abgetasteten Papier ein schwarzweißes oder farbiges Rasterbild und übergibt es an die Scan-Software via TWAIN-, ISIS- oder FAX-Schnittstelle. An dieser Stelle werden das Format und die Auflösung des Rasterbildes gewählt. Empfangene Faksimile-Dokumente unterscheiden sich von gescannten Dokumenten kaum, so dass sie in der Regel von der gleichen Software weiterverarbeitet werden.

Automatische Bildverarbeitung: Die Bilder werden wahlweise für die Qualitätsprüfung vorbereitet – sie werden entfleckt, leere Seiten entfernt, Helligkeit und Kontrast so eingestellt, dass eine optimale Lesbarkeit gewährleistet ist u. v. m.

Qualitätsprüfung: Der Scan-Operator führt, falls vorgesehen, eine Sichtkontrolle durch, greift manuell ein und wiederholt, falls notwendig, den Scan-Vorgang einzelner Seiten oder des ganzen Stapels. An dieser Stelle werden oft auch einfache Klassifikationsdaten wie beispielsweise die Stapelnummer manuell erfasst (Operator-Arbeitsplatz).

Texterkennung und Barcodes: Die vorverarbeiteten Bilder werden nun der Zeichenerkennung (OCR: Optical Character Recognition) zugeführt. Die Seiten werden in Leserichtung gedreht, der Text und die Barcodes werden erkannt und den Bildern zugeordnet.

Klassifikation: Die erkannten Texte und Barcodes können zur automatischen Klassifikation des Dokuments dienen. So werden beispielsweise Rechnungen, Lieferscheine und andere Transaktionsdokumente unterschieden oder die Steuererklärung der steuerpflichtigen Person zugeordnet. Ist eine automatische Klassifikation nicht oder nur teilweise möglich, wird diese Arbeit manuell durchgeführt (Index-Arbeitsplatz).

Erfassung von Metadaten (Indizierung): Informationen aus der manuellen Klassifizierung der erkannten Barcodes und weiteren Quellen werden zu Metadaten (Index-Daten) zusammengefügt und den Dokumenten zugeordnet.

Segmentierung und Kompression: Der Speicherplatzbedarf der gescannten, rohen Bilddaten ist sehr groß (45 MB für eine farbige A4-Seite mit 400 dpi). Durch leistungsfähige Kompressionsverfahren lässt sich die Datenmenge stark reduzieren (auf ca. 200 kB). Zudem ermöglicht ein spezielles Verfahren (MRC: Mixed Raster Content) eine weitere, signifikante Reduktion der Datenmenge (auf ca. 20 kB). Dazu ist allerdings eine Segmentierung – die Zerlegung des Bildes in Einzelteile wie Hintergrund, Text und Fotos – notwendig.

PDF/A-Erzeugung: Die verarbeiteten und komprimierten Bilder jeder Seite, der erkannte Text und die Metadaten werden zusammen mit der Farbcharakterisierung des Scanners (ICC-Farbprofil) zu einem PDF/A-Dokument zusammengefügt. Oft werden die Metadaten getrennt weiterverarbeitet (Index-Datei).

Digitale Signatur: Es kann eine digitale Signatur aufgebracht werden, um die rechtliche Nachvollziehbarkeit des Dokumentenzustands zum Zeitpunkt des Posteingangs sicherzustellen.

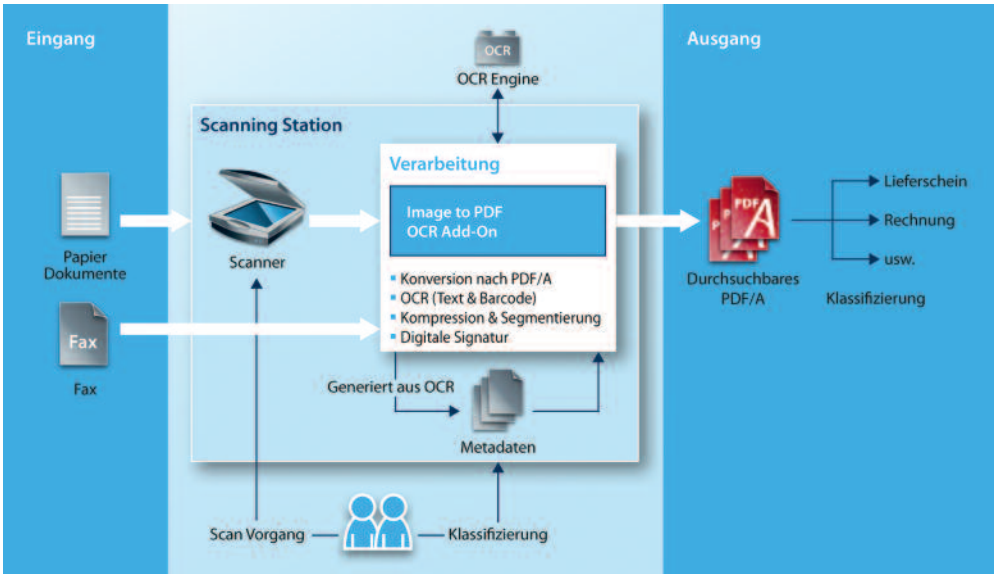
Validierung: Die PDF/A-Konformität des erstellten Dokumentes und die Gültigkeit der digitalen Signatur können überprüft und in einem Protokoll festgehalten werden.

Das Produkt der Digitalisierung: PDF/A

PDF/A ist ein ISO-Standard für die Verwendung des PDF-Formats in der Langzeitarchivierung elektronischer Dokumente. Er wurde erstmals am 1. Oktober 2005 als ISO-19005 veröffentlicht. Der PDF/A-Standard definiert „ein Dateiformat basierend auf PDF, genannt PDF/A, welches einen Mechanismus zur Verfügung stellt, um elektronische Dokumente so darzustellen, dass das visuelle Erscheinungsbild über die Zeit erhalten bleibt, unabhängig von den Werkzeugen und Systemen zur Herstellung, Speicherung und Reproduktion“. Der PDF/A-Standard ist kein neues Format, sondern definiert auf der Grundlage des PDF-Formats die Anforderungen an die Beschaffenheit der Dokumente, so dass sie für die verlässliche Langzeitarchivierung geeignet sind. In der Zwischenzeit sind auch Teil 2 und Teil 3 des Standards erschienen, damit das Format mit der Entwicklung Schritt halten kann.

Bietet das weit verbreitete TIFF Format nicht dieselben Eigenschaften? Auf den ersten Blick schon. Beide Formate können gescannte Rasterbilder speichern. PDF/A ist jedoch das modernere Format und bietet einige Vorteile. Die wichtigsten sind:

- leistungsfähige Kompressionsverfahren;
- standardisiertes Verfahren zur Volltextsuche;



- standardisierte digitale Signaturen (PAdES: PDF Advanced Electronic Signature) können zum Schutz der Integrität in das Dokument eingebettet werden;
- Metadaten werden in standardisierter Form (XMP: Extensible Metadata Platform) gespeichert und sind im Dokument eingebettet;
- PDF/A ist als universelles Dokumentenformat nicht nur für gescannte, sondern auch für digital erzeugte Dokumente bestens geeignet.

Architektur: Dezentral oder zentral?

Die Wahl der Architektur hängt stark von der Art, dem Umfang und der Regelmäßigkeit der Verarbeitung ab. Für den gelegentlichen, persönlichen Bedarf dient ein einfacher Multifunktions-Scanner mit eingebauter Scan-Software. Dafür müssen aber kaum Digitalisierungsprojekte gestartet werden. Die Frage ist eher, wie die unterschiedlichen Bedürfnisse in einem Unternehmen zu einer einheitlichen Scan-Strategie vereint werden sollen. Da gibt es die dezentralen, leistungsfähigen Multifunktionsgeräte (MFP) für den persönlichen Bedarf der Mitarbeiter, welche in jeder Abteilung stehen und die Scan-Straßen mit Stapel-Scannern in Dienstleistungszentren, die regelmäßige, hohe Dokumentenströme verarbeiten können.

In der Regel wird die für jedes Scan-Gerät spezialisierte Software dezentral betrieben, oft als Teil des Geräts selbst. Dies mag mit ein Grund für die wachsende Beliebtheit der Multifunktionsgeräte sein. Bei Hochleistungs-Scannern sind dezentrale Lösungen jedoch nicht so beliebt, weil sie teuer sind und die dezentrale Architektur die Prozesse verlangsamen können. Um diesen Problemen entgegenzuwirken, hat man begonnen, die aufwendigen und teuren Funktionen der Scan-Software zu zentralisieren. Dazu teilt man die Funktionen etwa wie folgt auf:

- Dezentral: Bildakquisition, Bildverarbeitung, Qualitätsprüfung, manuelle Klassifikation, Indizierung.
- Zentral: Zeichenerkennung, Segmentierung und Kompression, PDF/A-Erzeugung, digitale Signatur, Validierung.

Mit dieser Aufteilung erhöht sich die Skalierbarkeit der Architektur, was sich in geringeren Beschaffungs- und Betriebskosten sowie höheren Durchsätzen bei hohen Dokumentenvolumen auswirkt. *(www.pdf-tools.com)*

Beim Einsatz von Hochleistungs-Scannern werden Funktionen wie die Bildverarbeitung, Segmentierung und Kompression, PDF/A etc. vorzugsweise zentral durchgeführt.