PDF Tools: Hohe Bildqualität bei geringer Datenmenge

Vom Scan zur Information

Nadine Schuppisser von der PDF Tools AG: „Ein zentraler Scan-Server bietet sich als effiziente und vielseitige Lösung an, wenn große Volumen an gescannten Dokumenten aus verschiedenen Quellen verarbeitet werden sollen."

Mit einem zentralen Scan-Server-Dienst lassen sich große Mengen von Papierdokumenten elegant in elektronische Dokumente umwandeln, für die Weiterverarbeitung aufbereiten und im Langzeitarchiv ablegen. Ein Scan-Server, wie ihn die PDF Tools AG mit dem „3-Heights Scan to PDF Server" anbietet, wandelt gescannte Dateien und dazugehörige Indexdateien in das standardisierte Dateiformat PDF/A um.

Papier hat im Zeitalter von E-Rechnung, Online-Schalter und E-Commerce keineswegs ausgedient: Dokumente wie Rechnungen, Steuerformulare, Service-Berichte und Verträge werden nach wie vor oft auf Papier ausgefertigt, per Post verschickt und auf dem Postweg entgegengenommen. Spätestens innerhalb der Unternehmens- oder Behördengrenzen sind IT-Systeme für die Verarbeitung der Informationen zuständig – was auf dem Papier steht, muss eingescannt, in maschinenlesbarer Form aufbereitet, gespeichert und archiviert werden.

Gescannt wird typischerweise direkt in den Fachabteilungen mit Multifunktionsgeräten (MFP mit zusätzlicher Druck- und Faxfunktion) oder zentral mit Hochleistungs-Scannern. Scans fallen in den meisten Unternehmen an verschiedenen Stellen an: Am zentralen Eingang im Postbüro, an Scan-Stationen in den Abteilungen sowie auf Mobilgeräten, z. B. beim Kundenbesuch im Außendienst. Auch empfangene Faxmeldungen sind nichts anderes als gescannte Bildinformationen.

Vom Bild zum standardisierten Dokument

Beim Scannen entsteht zunächst immer ein Faksimile in Form einer Bilddatei. Dabei kommen Rasterformate wie TIFF und JPEG zum Einsatz. Ein Rasterdokument ist jedoch bloß ein Abbild ohne Zusatzinformationen. Texte sowie in Barcodes enthaltene Informationen müssen nach dem Scannen durch Texterkennung (OCR, Optical Character Recognition) aus dem Bild extrahiert werden. Idealerweise werden Text und Bilddarstellung danach gemeinsam im gleichen Dokument gespeichert. Dies vereinfacht die Ablage und sichert sowohl das Erscheinungsbild als auch den Informationsgehalt des Ursprungsdokuments.

Als Format für die standardisierte Ablage und für die Langzeitarchivierung gescannter wie auch elektronisch erzeugter Dokumente hat sich PDF/A etabliert. Der PDF/A-Standard unterstützt die gewünschte Speicherung von Bild- und Textinformationen im gleichen Dokument. Die Dokumente sind damit per Volltextsuche durchsuchbar.

Für die Bildinformationen arbeitet PDF/A mit leistungsfähigen Kompressionsverfahren. Dadurch verringert sich die ursprüngliche Dateigröße ohne Informationsverlust massiv. Dies fällt besonders ins Gewicht, wenn neben Schwarzweiß- auch Farbbilder enthalten sind und die Farbinformationen für die weitere Nutzung erhalten werden sollen.

Zusätzlich erlaubt PDF/A, Metadaten wie beispielsweise Klassifizierungsinformationen direkt im Dokument zu speichern – hierbei kommt das XMP-Format (Extensible Metadata Platform) zum Zug, das wie PDF/A als eigener ISO-Standard definiert ist. Eine weitere Möglichkeit von PDF/A ist die digitale Signierung, um die Authentizität der Dokumente und die Integrität der Inhalte zu gewährleisten. Insgesamt bietet PDF/A die Sicherheit eines internationalen, funktional umfassenden und auf langfristige Stabilität ausgerichteten Dokumentenstandards.

Dezentral scannen, zentral verarbeiten

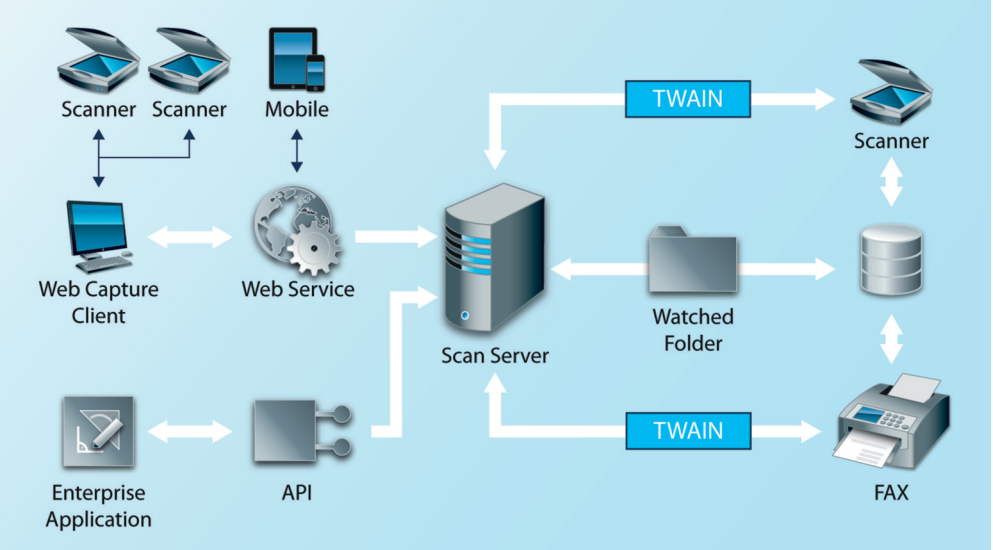
Das eigentliche Scannen stellt keine hohen Leistungsanforderungen an die Hardware und Software. Im Prinzip lassen sich „Scans" bereits mit einer einfachen Digitalkamera erzeugen. Die darauf folgenden Bearbeitungsschritte verlangen deutlich mehr an Rechenleistung und Intelligenz. Bildkompression, OCR und Konversion zu PDF/A sind relativ aufwändige Vorgänge. Zumal es dabei zwei gegenläufige Bedürfnisse zu berücksichtigen gilt: Die zuverlässige Texterkennung setzt eine möglichst hohe Bildqualität voraus.

PDF Tools

Die PDF Tools AG ist ein Hersteller von Software-Lösungen und Programmierkomponenten für die PDF- und PDF/A-Erzeugung, Bearbeitung, Wiedergabe und Archivierung.
(www.pdf-tools.com)

Damit steigt der Speicherbedarf. Für die Ablage wünscht man dagegen eine möglichst geringe Datenmenge. Software, die beide Ansprüche in Einklang bringen soll, stellt hohe Ansprüche an die Rechenleistung – vor allem dann, wenn ein großes Volumen an gescannten Dokumenten verarbeitet werden soll.

Dazu kommt ein weiterer Aspekt: Für die Einbettung von Index-, Klassifizierungs- und anderen Metadaten sowie digitalen Signaturen werden oft Informationen von anderen Arbeitsstationen und aus verschiedenen IT-Systemen benötigt. Diese dezentral vorhandenen Daten müssen für die Erstellung des PDF/A-Dokuments zusammengeführt werden.



Der Scan-Server arbeitet als zentrale PDF/A-Aufbereitungsinstanz und verarbeitet Scans aus verschiedenen Quellen.

Leistungsintensive Operationen

Die Lösung für beide Probleme ist ein zentraler Scan-Server – ein Beispiel ist der „3-Heights Scan to PDF Server“ aus dem Hause PDF Tools AG. Er nimmt die eingescannten Bilddateien entgegen, analysiert die Dokumente und erzeugt ein PDF/A-Dokument mit allen Text- und Bildinformationen in jeweils passender Kompression. Optional markiert er das Dokument mit einem Zeitstempel oder einer digitalen Signatur. Die erfassten Informationen stehen damit sowohl für menschliche Leser als auch zur automatisierten Weiterverarbeitung mit IT-Anwendungen in standardisierter, qualitativ hochwertiger Form zur Verfügung.

Ein zentraler Scan-Server vereinfacht zusätzlich die Software-Verteilung und Wartung. Auf den Scan-Stationen muss keine umfassende Scan-Software mit integrierter OCR-Funktionalität einzeln ausgerollt, konfiguriert und gepflegt werden. Eine elementare Operator-Anwendung zur Bildakquisition genügt. Probleme bei den komplexeren Verarbeitungsschritten müssen nicht individuell an den Arbeitsplätzen gelöst werden. Stattdessen wird der Scan-Server-Dienst auf einer Testinfrastruktur implementiert, wo sich zunächst alle Probleme analysieren und Fehler beheben lassen. Danach wird der Dienst in den produktiven Betrieb überführt.

Damit sich der Scan-Server optimal an die jeweilige Umgebung anpassen und bei Bedarf durch Aufteilung auf verschiedene Rechner skalieren lässt, sind die Aufgaben beim „3-Heights Scan to PDF Server“ auf mehrere Subsysteme verteilt:

- Der eigentliche Scan-Server nimmt Aufträge für die Konversion ins PDF/A-Format entgegen, delegiert die Texterkennung an den OCR-Server und kombiniert die OCR-Resultate, das gescannte Bild und die Metadaten zum fertigen PDF/A-Dokument.
- Der OCR-Server nimmt vom Scan-Server Aufträge zur Erkennung von Texten und Barcodes entgegen, bereitet die Bildinformationen durch Operationen wie Geraderichten und Entfernen von Störungen für die bestmögliche Texterkennung auf, gliedert das Dokument in Text-, Barcode- und Bildbereiche und führt die Erkennung durch.

Für dezentral erzeugte Scans bietet der Server zwei zusätzliche Dienste: Ein Watched-Folder-Service übermittelt alle Dateien, die in bestimmten Verzeichnissen abgelegt wurden, zur automatischen Weiterverarbeitung an den Scan-Server. Mithilfe eines Web-Service nimmt der Scan-Server Aufträge entgegen, die über eine webbasierte Anwendung erfasst wurden, und schickt die umgewandelten Dokumente an den Auftraggeber zurück. Der Scan-Server kann darüber hinaus weitere nützliche Aufgaben übernehmen, darunter die Validierung der erzeugten PDF/A-Dokumente auf Konformität mit dem ISO-Standard, das Markieren der Dokumente mit einem Wasserzeichen und die Kombination verschiedener Einzeldokumente, die zum gleichen Geschäftsfall gehören, zu einem Gesamtdokument.

Ein zentraler Scan-Server bietet sich als effiziente und vielseitige Lösung an, wenn große Volumen an gescannten Dokumenten aus verschiedenen Quellen verarbeitet werden sollen. Er wandelt die gescannten Bilddaten in standardisierte, durchsuchbare PDF/A-Dokumente mit reichem Informationsgehalt auf, entlastet die Scan-Stationen von leistungsintensiven Verarbeitungsschritten, unterstützt die Einbindung weiterer IT-Systeme und hilft, einen unternehmensweit einheitlichen Dokumentenstandard einzuhalten.