## 3.3 Sicher archivieren mit PDF/A

Unternehmen sind mit unzähligen Dokumenten konfrontiert, wie beispielsweise Office- oder CAD-Dateien oder PDFs in unterschiedlichen Varianten. Werden diese in ihrem ursprünglichen Format aufbewahrt, so ist die Gefahr groß, dass sie nach Jahren nicht mehr lesbar sind. Daher ist für die Archivierung eine konsequente Konvertierung in ein stabiles Format notwendig, das die langfristige Reproduzierbar- und Lesbarkeit sicherstellt.

PDF/A als ISO-Standard ist hierbei das Format erster Wahl. Es bietet entscheidende Vorteile gegenüber anderen Formaten, wie beispielsweise TIFF. Dazu zählen kleinere Dateigrößen, die Fähigkeit zur Volltextsuche und das einfache Handling. Nicht zu vergessen ist die Tatsache, dass jede PDF/A-Datei immer eine PDF-Datei ist. Für kaum ein anderes Format gibt es derart zahlreiche Werkzeuge und Lösungen. Ziel ist, dass die Konvertierung der kaum beeinflussbaren Formatvielfalt nach PDF/A automatisiert erfolgt, was nicht ohne Weiteres möglich ist. PDF/A ist der ISO-Standard 19005 für die Langzeitarchivierung im PDF-Format. Es ist zwar nicht vorgeschrieben, PDF/A zu nutzen, er hat aber eine allgemeine und breite Akzeptanz gefunden. Der Standard bewertet und regelt, welche PDF-Funktionen in puncto Archivierung sicher sind. Diese Vorschriften garantieren eine langfristige Lesbarkeit der Dokumente- und zwar unabhängig davon, mit welcher Anwendungssoftware und auf welchem Betriebssystem sie ursprünglich erstellt wurden.

### Drei Szenarien für die PDF/A-Konvertierung

Für die Wandlung von Dokumenten nach PDF/A sind architektonisch grundsätzlich drei Szenarien möglich:

**Client-seitige Konvertierung:** Hierbei startet der Anwender die Konvertierungs-Engine und korrigiert eventuell auftretende Fehler. Die Übergabe an das Archiv erfolgt manuell oder automatisiert über die Software. Da der Konvertierungprozess sehr rechenintensiv ist, sind damit schwache Desktop-Rechner zu einem hohen Grad ausgelastet. Anwender müssen in der Regel warten, bis die Wandlung nach PDF/A abgeschlossen ist. Ein weiterer Nachteil der Client-seitigen Konvertierung ist, dass die PDF/A-Dateien verteilt erzeugt werden und somit deren tatsächliche Konformität nur schwer kontrollierbar ist. Eine Lösung wäre eine nachgelagerte Validierung auf dem Server, wodurch wiederum ein Overhead entsteht.

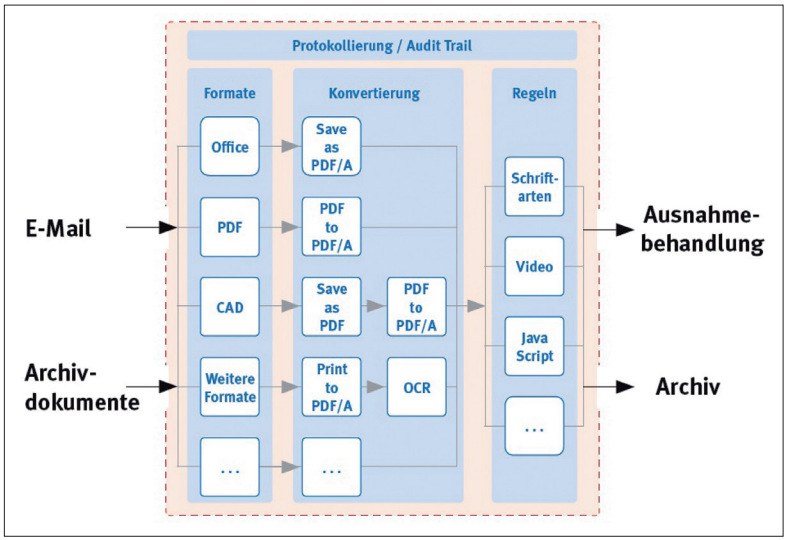
**Mischform:** Alternativ wählt der Anwender über den Client die zu archivierenden Dokumente aus und lädt sie dann zur Konvertierung auf den Server. Auch hier kann eine anschließende Qualitätssicherung durch den Anwender erfolgen. Da der Konvertierungsprozess sehr komplex und damit zeitaufwendig ist, sollten Vorkehrungen für diesen asynchronen Ablauf getroffen werden, sodass der Anwender jeweils eine Rückmeldung über die Konvertierung erhält.

**Serverseitige Konvertierung:** Bei diesem Szenario erfolgt die Konvertierung ohne Anwenderinteraktion. Dafür müssen sämtliche notwendigen Funktionen auf dem

Server hinterlegt sein und entsprechend verwaltet beziehungsweise gesteuert werden. Ist ein hohes Dokumentenvolumen zu verarbeiten, empfiehlt es sich, die Konvertierung nach PDF/A auf dem Server automatisiert vorzunehmen.

### 3.3.1 Wege aus dem Format-Chaos

Wie bereits oben beschrieben, können die zu archivierenden Dateiformate auch schon in einem Unternehmen sehr unterschiedlich sein: Dateien aus beliebigen Office-Dokumenten, PDFs in unterschiedlichen Nuancen, gescannte Dokumente, CAD-Dateien, branchenspezifische Formate oder vielleicht sogar PDF/A-Dateien, bei denen aber dann noch zu prüfen wäre, ob sie wirklich dem ISO-Standard entsprechen. Es ist also ein „zu bändigender Format-Zoo", mit dem Unternehmen konfrontiert sind. Da Unternehmen zahlreiche Dateien von externen Stellen erhalten, helfen Richtlinien, die vorschreiben, welche Dateiformate für die Konvertierung nach PDF/A anzuliefern sind, nur eingeschränkt. Es ist vielmehr notwendig, abhängig vom Ursprungsformat zu definieren, welche Konvertierungstechnik angewandt und welche Tools eingesetzt werden sollen. Die schlechte Nachricht: Ein generischer Konverter, der alle möglichen Formate nach PDF/A konvertiert, existiert nicht. Die gute Nachricht: Es stehen zahlreiche Tools zur Verfügung, die, setzt man sie für ihren jeweiligen Anwendungsfall ein, 95 Prozent oder sogar bis zu 99 Prozent der Quellformate automatisch und zuverlässig nach PDF/A konvertieren.




**Umfangreich:** Bei einer „Everything to PDF/A-Lösung" werden alle Konvertierungstechniken benötigt. (Quelle: LuraTech Europe GmbH)

Damit überhaupt eine Wandlung nach PDF/A möglich ist, müssen die unterschiedlichen Dateiformate lesbar sein. Somit ist es erforderlich, dass die Client-Version der Anwendungssoftware auf dem Server installiert ist. Das mag vielleicht zunächst trivial klingen, ist es aber nicht. Produziert beispielsweise MS Word bei der Ausführung des Befehls „Speichern unter PDF/A" eine Fehlermeldung, muss vermieden werden, dass daraufhin der gesamte Server blockiert ist. Über entsprechende Regeln wird definiert, wie die Lösung in diesem Fall reagieren soll. Zusammengefasst besteht also die Aufgabenstellung darin, die verschiedenen Anwendungsapplikationen und Konvertierungswerkzeuge zu einer stabilen Gesamtlösung zu integrieren. Und wenn die Dateien schon angefasst und verarbeitet werden, dann bietet es sich an, andere Aufgaben gleich mit zu erledigen. Dazu gehören beispielsweise das Skalieren der Auflösung bei gescannten Dokumenten, die OCR-Erkennung oder die Erzeugung von Derivaten. Die dazu notwendigen Produkte müssen ebenfalls in die Gesamtlösung eingebunden werden, die dann sämtliche Funktionen in einer Serverlandschaft integriert. Zusätzlich müssen die Prozesse so aufgesetzt sein, dass sie möglichst automatisiert ablaufen und über ein Eskalationsmanagement eventuelle Fehlermeldungen abfangen.

### 3.3.2 Wenn die PDF/A-Konvertierung nicht klappt

Doch damit ist es leider nicht getan: Es ist durchaus wahrscheinlich, dass sich nicht jedes Quellformat automatisiert nach PDF/A konvertieren lässt. Die Erfahrung aus zahlreichen Projekten zeigt, dass dies am häufigsten auftritt, wenn PDF-Dateien auf einer Ebene mit einem Passwortschutz versehen sind.

Des Weiteren kann eine schlechte Qualität der Ursprungsdatei die Konvertierung verhindern. Typische Beispiele sind fehlende Schriften oder nicht zulässige Inhalte, wie Videos oder JavaScripts. Zu einem verschwindend kleinen Teil tauchen Probleme mit Office-Dokumenten auf, wenn beispielsweise eine PowerPoint-Datei mit Transparenzen in PDF/A-1 zu wandeln ist. Um den gesamten Prozess nicht zu unterbrechen, muss auch hier definiert sein, wie das System verfahren soll. So kann man hinterlegen, dass bestimmte nicht vorhandene Schriftarten durch andere ersetzt werden. Unerlaubte Inhalte in einem Dokument, wie etwa Videos, könnten zum Beispiel vor der Konvertierung nach PDF/A durch ein Bild ersetzt werden, und das Video könnte als Originaldatei abgespeichert werden.

Je mehr Regeln hinterlegt sind, desto stärker nähert man sich der 100-prozentigen Dunkelverarbeitung. Tauchen dennoch Dokumente auf, die unter Berücksichtigung aller Regeln nicht nach PDF/A konvertierbar sind, kann beispielsweise der Administrator hinzugezogen oder das Dokument an den Absender zurückgesendet werden. Die Erfahrung zeigt, dass die Lernkurve bei einem solchen Projekt schnell nach oben geht und die Zahl der Dateien, die automatisiert nach PDF/A konvertiert werden, kontinuierlich steigt. Welche Verarbeitungsschritte die Dokumente durchlaufen, muss aus Compliance-Gründen protokolliert werden.

### 3.3.3 Bestandteile einer PDF/A-Lösung

Eine Fragestellung, die bei Archivierungsprojekten immer wieder auftaucht, ist, ob eine visuelle Qualitätskontrolle nach der PDF/A-Konvertierung erfolgen soll. Insbesondere bei einer hohen Dokumentenzahl steht der Aufwand aber in keinem Verhältnis zum Nutzen und wird deshalb in der Regel nicht umgesetzt.

Folgende Bestandteile zählen zu einer „Everything to PDF/A"-Lösung:

- Herstellung der Betriebsstabilität
- Abbilden von Konvertierungsstrategien
- Entpacken von ZIP-Dateien und Ähnlichem
- Zuordnung des Eingangsmaterials zum richtigen PDF/A-Verarbeitungsschritt
- Konvertierung beliebiger „Born Digital"-Dokumente (Office, CAD usw.)
- Umwandlung von PDF nach PDF/A
- Komprimierung und Wandlung gescannter Dokumente, ggf. Durchführung von OCR
- Validierung von PDF/A-Dateien
- Abfangen und Eskalieren bei Fehlern und nicht geregelten Ausnahmen
- Bedienung der Schnittstellen zum DMS-/Archiv-System
- Ggf. Zusatzfunktionen wie Skalieren, Derivate erzeugen
- Protokollierung jedes Verarbeitungsschrittes (zum Beispiel nach GoBS)

### 3.3.4 Fazit

Eine „Everything to PDF/A"-Lösung ist weit mehr als ein Konvertierungs-Tool. Sie ist vor allem eine skalierbare „Managementsoftware", die eine hohe Betriebsstabilität sicherstellt, Konvertierungsstrategien sowie „Reparaturregeln" abbildet und darüber hinaus Fehler oder nicht geregelte Ausnahmen abfängt und eskaliert. Ohne einen ausgewiesenen PDF/A-Experten ist sie aufgrund ihrer Komplexität kaum realisierbar. Weitere Informationen über das PDF/A-Format stellt die PDF/A Association auf ihrer Website pdfa.org zur Verfügung. Der internationale Interessenverband verfolgt das Ziel, PDF-Anwendungen für digitale Dokumente zu fördern, die auf offenen Standards basieren. Dazu setzt sich der Verband für eine aktive Wissensvermittlung und den Austausch von Know-how und Erfahrungen für alle Interessengruppen weltweit ein.

Carsten Heiermann

**Carsten Heiermann** ist Geschäftsführer der LuraTech Europe GmbH.

*Dieser Artikel basiert auf einem Beitrag unserer Schwesterpublikation Computerwoche.*