



# zeitungstechnik

Vom Folianten zur Volltextsuche

## PPS wandelt historische Zeitungsbände ins PDF-Format

Klaus von Prümmer

Sonderdruck aus „ifra-zeitungstechnik“  
September 2004

Vom Folianten zur Volltextsuche

# PPS wandelt historische Zeitungsbände ins PDF-Format

**Seit praktisch alle Zeitungen ihre Seiten vollständig elektronisch produzieren, hat es sich durchgesetzt, im Archiv neben den gebundenen Exemplaren und dem herkömmlichen Mikrofilm auch komplette Jahrgänge als PDF-Dateien zu konservieren. In der Praxis ist dies längst das bevorzugte Medium für all jene, die auf unlängst erschienene Artikel oder Anzeigen zugreifen wollen, denn PDF erlaubt die bequeme Suche im Volltext, und die gefundenen Seiten können gewissermaßen 1:1 an den Arbeitsplatz geholt werden.**

Je weiter sich dieses Verfahren durchsetzt, umso hinderlicher erscheint es, wenn man historische Archivbestände nur in der herkömmlichen Form nutzen kann: Man begibt sich ins Archiv und sucht alte Mappen und Bände durch. Bei Recherchen ohne präzise Anhaltspunkte kann das sehr zeitaufwendig sein.

Wie man dieses Dilemma beseitigen kann, liegt seit langem auf der Hand: Die Umwandlung der Altbestände in digitale Dateien, die dann mit einem Volltext-Index erschlossen werden können. Schon vor dreißig Jahren hat eine amerikanische Zeitung ihre Altbestände in Asien abtippen lassen. Inzwischen weiß man, dass ein Zeitungsarchiv aber mehr ist als nur die Verwaltung von Rohdaten. Auch die Aufmachung eines Beitrags und sein redaktionelles Umfeld gehören dazu, wie überhaupt das Layout einer Publikation maßgeblich ihr Image auf den Inhalt überträgt: Es macht einen Unterschied, ob eine Nachricht im Wall Street Journal oder in einer Boulevard-Zeitung erschienen ist.

Es hat allerdings lange gedauert, bis Hardware und Software überhaupt in der Lage waren, Zeitungssseiten so auf einem Bildschirm darzustellen, dass man darin tatsächlich lesen konnte. Was heute für jeden PC selbstverständlich ist, war vor zehn Jahren schlichtweg noch undenkbar. Der Fortschritt ist nicht nur immer schnelleren Computern, größeren Festplatten und besseren Bildschirmen zu verdanken, sondern ebenso dem universellen Portable Document Format von Adobe, kurz PDF, das

heute als weltweiter Standard nicht nur auf PCs mit Windows, sondern auf allen zeitgemäßen Computer-Plattformen einsetzbar ist.

## Eine Vision wird Wirklichkeit

Die Vision, historische Zeitungsbände in PDF zu konvertieren, ist also nahe liegend. Die Realisierung war aber keineswegs einfach. Das liegt in erster Linie an der Natur der Zeitungsbände in den Archiven. Viele Jahrgänge sind in einem Zustand, der sie schon für Menschen kaum noch lesbar erscheinen lässt. Außerdem sind die dicken Folianten reichlich sperrig – mit herkömmlichen Scannern kommt man ihnen einfach nicht bei. Schon vom Mikrofilm her kennt man das Phänomen, dass die Seiten im Bund eines aufgeschlagenen Bandes gewölbt sind und deshalb der Text verzerrt dargestellt wird. Ein Scannen in dieser Form würde niemals zu einem Ergebnis führen, das für eine optische Texterkennung als Grundlage für einen Volltext-Index ausreicht.

Andererseits ist der beklagenswerte Zustand vieler Archivbände selbst ein wesentlicher Grund, die Digitalisierung dieser Materialien zügig in Angriff zu nehmen, bevor sie endgültig zerfallen. Viele Archivre lassen schon heute keine Fremden mehr an ihre wertvollen Bestände.

All diese Schwierigkeiten erforderten einen Mann der Praxis. Siegfried Peis, den Geschäftsführer der PPS Prepress Systeme GmbH ([www.prepress-systeme.de](http://www.prepress-systeme.de)) in Bad Homburg, D, einen gelernten Setzer und Druckingenieur, der auch Erfahrungen in Druck, Buchbinderei und der vielgestaltigen Produktion einer Werbeagentur gesammelt hat, ließ die Herausforderung nicht ruhen. Er suchte sich handelsübliche Gerätschaften zum Scannen von großformatigen Vorlagen, Software für die Bildverbesserung, die Layout-Analyse und die Texterkennung unter erschwerten Bedingungen. Um das Ziel der fließbandmäßigen Abarbeitung von Zeitungsbänden zu erreichen, mussten all diese Einzelkomponenten noch angepasst werden. Doch den entscheidenden



In wenigen Sekunden fährt der großformatige Scanner über die Doppelseite. Entscheidend ist ein mechanischer Zusatz zu der Buchwippe, der die Zeitungssseiten völlig plan hält, ohne den Band zu beschädigen. Für höchste Scann-Genauigkeit wurde der massive Metallaufbau mikrometergenau justiert.



den Schritt leistete Peis mit seiner Mannschaft selbst: Sie fanden eine rein mechanische, aber höchst wirkungsvolle Lösung, den Zeitungsbund auf einer Buchwippe so einzuspannen, dass zwei nebeneinander liegende Seiten vollkommen plan liegen. Die mechanische Beanspruchung des Bandes ist dabei so gering wie möglich: „Noch nie“, sagt Peis stolz, „ist bei uns ein Band beschädigt worden.“

Mit diesem inzwischen zum Patent angemeldeten Bauteil entstand eine einsetzbare Anlage, die zwar noch den Charakter eines Vorserienmodells hat, aber höchst produktiv arbeitet. Der Fachmann am Scanner liest die Doppelseiten ein und legt die Rohdaten als TIFF-Dateien auf einem Server ab. Mit dem Programm Preview werden die Seitenbilder elektronisch gereinigt, bevor mit Abby-OCR zusätzlich zu dem Bild eine lesbare Textdatei erzeugt wird. Schließlich wird jede Seite am Bildschirm optisch kontrolliert. Die inzwischen eingespielte Mannschaft schafft in einer Schicht mehr als zweitausend Seiten. Für die Volltext-Erschließung und die Recherche bietet PPS ein so genanntes Knowledge Management Tool auf der Basis von Converas Sesam.Ware an. Diese Retrieval-Software ist bei OCR-Lese Fehlern ungemein tolerant – ein wichtiger Aspekt, da die Texterkennung bei teilweise zerstörten Vorlagen zwangsläufig nicht immer perfekt sein kann.

#### 46 Jahrgänge auf zwei Festplatten

Die beiden Kunden, die mit PPS zusammen arbeiteten und mithalfen, alle Schwierigkeiten zu erkennen und aus dem Weg zu räumen, waren Die Zeit ([www.zeit.de/archiv/index](http://www.zeit.de/archiv/index)) in Hamburg und die Leipziger Volkszeitung ([www.lvz-online.de](http://www.lvz-online.de)). Die Bestände beider Publikationen reichen bis in die unmittelbare Nachkriegszeit zurück. Inzwischen sind die 145 811 Seiten der Zeit-Jahrgänge von 1946 bis 2002 komplett digitalisiert. Als PDF-Dateien beanspruchen sie gerade einmal 400 Gigabyte Speicherplatz – das passt heutzutage auf zwei Festplatten eines Desktop-PCs. Die Kosten für Speicher und Datensicherung spielen da keine Rolle mehr. Komplexer ist das Retrieval, da die Zeit nicht das Standard-Modul von PPS nutzt, sondern die historischen Daten vollkommen in das Redaktionssystem des Hauses integrieren will. Rüdiger Pflughaupt von der Gruppen-



Links: Höchste Zeit, wenigstens den Inhalt zu retten. Das Papier der Nachkriegsjahre zerfällt unauffaltam. Spezialisten sind in der Lage, auch solche Bände mit brauchbarem Ergebnis zu digitalisieren. Rechts: Siegfried Peis begutachtet die historische Titelseite der Zeit. Trotz der Unzulänglichkeiten des vom Texterkennungsprogramm erstellten Volltexts ist dieser für Recherchen verwertbar. Einzelne Artikel lassen sich mit simplem Cut & Paste in ein Textverarbeitungsprogramm übernehmen.



leitung IT sieht darin eine besonders schnelle und komfortable Suchmöglichkeit für die Redakteure der Zeit, die dann bei der gleichzeitigen Nutzung von PDF-Quellen und aktuellem Material nicht umdenken müssen und sich ganz auf ihre journalistische Arbeit konzentrieren können.

Inzwischen sieht der Zeit-Projekt-leiter Frank Rödel die wesentlichen Ziele schon „zu achtzig Prozent“ erreicht. Es sei darum gegangen, angesichts der zerbröselnden Bestände des Papierarchivs wenigstens die Inhalte zu retten und die Bestände vor 1995 für die elektronische Recherche zugänglich zu machen – ab dann hat man die Daten in der konzerneigenen Datenbank von Genios. Da PDF aber viel mehr bietet als nur Rohtext, wurden alle Jahrgänge bis zu dem Zeitpunkt erfasst, ab dem PDF-Seiten schon im normalen Herstellungsprozess erzeugt wurden.

Rödel ließ offen, ob man die PDF-Seiten auch öffentlich zugänglich machen werde. Angesichts des Verkaufserfolgs mit der CD-ROM-Ausgabe der Zeit seit 1995 sieht Rödel durchaus Vermarktungsmöglichkeiten. Immerhin seien die Voraussetzungen jetzt dafür gegeben, aber organisatorische und rechtliche Fragen müssten noch geklärt werden. Interessenten könnten frisch produzierte „Sonderdrucke“ erhalten, der Leserservice könne nun mit dem PDF-Material vielfältige Wünsche erfüllen. Sehr gefragt seien bei der Zeit themenbezogene Materialien für den Schulunterricht.

Derzeit beraten Peis und Rödel, auch über die Digitalisierung des Zeit-Magazins. Die spezielle Anforderung, ein Kultobjekt gattungsgerecht abzubilden, hat zusätzlichen Reiz, aber auch Anforderungen, die bei der Massenverarbeitung von Zeitungs-

seiten eher nachrangig erscheinen. Mit größerem Nachdruck arbeitet PPS daran, dass seine OCR-Software auch mit der Frakturschrift zu brauchbaren Ergebnissen kommt, damit auch Zeitungen vor der Umstellung in den vierziger Jahren digitalisiert werden können. Die Fortschritte sind durchaus ermutigend.

PPS glaubt, mit seiner Lösung nun im großen Stil Zeitungsarchive digitalisieren zu können. Auch ausländische Publikationen haben sich angekündigt, vor allem aus Skandinavien, der Heimat des Scoop-Redaktionssystems, das PPS in Deutschland vertreibt. Peis beabsichtigt, mehrere Arbeitsplätze für das Scannen einzurichten und diese in zwei Schichten zu besetzen. Eine solche Mannschaft könnte dann in kürzester Zeit ein komplettes Zeitungsarchiv digitalisieren.

Klaus von Prümmer ([pruemmer@ifra.com](mailto:pruemmer@ifra.com)) ist Manager

R I V C O M

WIE ZEITUNGEN  
DIE VORTEILE  
VON XML  
NUTZEN KÖNNEN

beratung • implementierung • schulung

www.rivcom.com  
GB: +44 (0)1793 792000  
USA: +1 (212) 222 4332  
IfraExpo 2004: Stand 1145