Sie betreuen ein Digitalisierungsprojekt?

Wir erlauben uns, Ihnen Tips aus der Erfahrung von elf Jahren Digitalisierung und der Produktion von über zehn Millionen Zeitungsseiten weiterzugeben.

Am Anfang steht das Archiv – ein riesiger Haufen Papier in großen Räumen mit Archivschränken. Manchmal befindet sich das Archiv aber auch in mehr oder weniger geeigneten Kellerräumen, nicht immer ideal für die Aufbewahrung der wertvollen Archivarien. Hier schlummern die Zeitungen der letzten 100 oder mehr Jahre, jahrgangs-, quartals- oder monatsweise zu Büchern gebunden, und warten darauf, zum Leben erweckt zu werden. Diese Bücher werden nicht ewig halten und mit deren Verlust derselben gehen vor allem die in ihnen enthaltenen Informationen verlorengehen. Informationen, die das wertvollste Gut einer Zeitung darstellen.

Im heutigen digitalen Zeitalter bietet der Zugang zu und die Recherche in Archiven einen zusätzlichen Anreiz zur Nutzung des Onlineangebots. Heute ist dieser Markt noch ein kleines Pflänzchen. Aber es wird sich zu einem Baum entwickeln, wenn die Möglichkeiten der Nutzung erst einmal erkannt sind. Es gibt Millionen Leser, die gerne in Archiven recherchieren würden und auch bereit sind, dafür Geld auszugeben.

Der Nutzen eines digitalen Archivs wird bei einigem Nachdenken schnell klar. Abgesehen von den Kundenbindungen, über den Vertrieb des zeitungseigenen Informationsmaterials und der Bestandssicherung, wird man viele Möglichkeiten der Verwertung und Refinanzierung finden.

Der erste Schritt zur Digitalisierung heißt nicht wollen, sondern tun – und man beginnt mit der Bestandsaufnahme.

Die Digitalisierung eines Archivs ist für jeden Verlag eine einmalige Angelegenheit. Ausschlaggebend für den Erfolg eines digitalen Archivs ist die Qualität der Ergebnisse und nicht der Preis. Die komplexe Aufgabe der Digitalisierung eines Zeitungsarchivs vom Scan bis zum automatisch separierten Einzelartikel erfordert jahrelange Erfahrung und beginnt mit der buchbinderischen Vorbereitung der zu digitalisierenden Bestände.

Die Wahl der Qual hat der Laie bei der Frage, wie scannen wir die Bestände. Graustufe, Bitonal oder Farbe? Die Beantwortung dieser Frage ist einfach und wird durch die Vorlage bestimmt.

Graustufe wird gewählt für Fotos, Zeitschriften die in Graustufe gedruckt sind (Tiefdruck).

Bitonal werden alle Tageszeitungen gescannt, die auch in s/w gedruckt sind, Zeitungen im Buchdruck oder im Offsetdruck. In diesen Druckverfahren gibt es keine Graustufe. Zu beachten ist dabei beim Scannen, dass die Scan-Einstellungen so gewählt sind, dass die Rasterpunkte korrekt dargestellt und auch die kleinsten Punkte erhalten bleiben. Eine Einstellung der Scanauflösung von 300 dpi bringt für die spätere Texterkennung die besten Ergebnisse.

Farbe wird natürlich bei farbigen Vorlagen angewandt. Auch hier ist eine Auflösung von 300 dpi für die spätere Texterkennung ausreichend.

Die Scanqualität ist maßgebend für alle weiteren Bearbeitungsschritte bis zur Separierung der Einzelartikel.

Die Verarbeitung nach dem Scannen

Geraderichten und Säubern. Die bereits beim Scannen getrennten Seiten werden nun für die OCR vorbereitet. Dazu müssen sie exakt geradegerichtet sein, Schmutzpartikel zwischen den Zeilen müssen entfernt, weiße Punkte in großen Überschriften müssen gefüllt werden. Bei größerer Anzahl weißer Punkte in Überschriften kann es passieren, dass die OCR solche Schriften als Bild erkennt und nicht als Text darstellt. Das heißt, diese würden dann im Text fehlen. Während Schmutzpartikel entfernt werden, sind Bilder und Graustufen geschützt, damit diese bei der Säuberung nicht verändert werden.

Text- und Layouterkennung. Die Layouterkennung der OCR-Software ist nicht ausreichend für die spätere Artikelseparierung. Daher ist es erforderlich, nach dem OCR-Lauf das Layout zu korrigieren. Durchläufe, in denen die Zeilen von mehreren nebeneinanderliegenden Spalten zusammengeführt werden, müssen korrigiert und die Spalten korrekt erkannt werden. Ein zweiter Korrekturlauf erkennt die Artikelelemente wie Dachzeile, Titel, Untertitel, Vorspann, Text, Bild und Bildunterschrift und versieht diese Elemente mit der entsprechenden Kennung in der XML-Datei. Ein nochmaliger abschließender Lauf mit der OCR-Software, bei dem es nur noch um die Zeichenerkennung geht, liefert die gewünschten Ergebnisse.

Der letzte Punkt im Workflow der Zeitungsdigitalisierung ist wohl der komplizierteste: die Automatische Artikel-Separierung.

AA-S steht für Automatische Artikel-Separierung.
Ziel der AA-S ist die automatische Erstellung von Einzelartikeln ohne jeden manuellen Eingriff. Die AA-S ordnet die erkannten Artikelelemente und stellt den korrekten Lesefluss her. In der vor-digitalen Produktion mit „kunstvollem“ Schachtelumbruch ist dies äußerst kompliziert. Es gelingt nicht immer, aber mit einer Erkennungsrate von deutlich über 80% ist dies schon eine beachtliche Rate. Nur eine automatisierte Arbeitsweise ermöglicht die Verarbeitung dieser gigantischen Datenmengen. Im Durchschnitt werden aus einer redaktionellen Seite etwa 10 Einzelartikel extrahiert, so dass schnell Artikel in Millionenzahl zu verarbeiten sind.

Artikelelemente, die wir erkennen:

- Dachzeile
- Titel
- Untertitel
- Zwischentitel
- Vorspann
- Text
- Bildunterschrift
- Bild
- Autoren
- Ressorts
- Rubriken
- Artikelfortsetzungen

Artikelsorten, die wir erkennen:

- Redaktionelle Artikel -> 98 %
- Werbung -> 65 % (wird gefiltert)
- Todesanzeigen -> 87 %
- Bilder + Bildunterschrift -> 90 %

Durchschnittliche Genauigkeit der Erkennung bei Tageszeitungen:

- Lesefluss des Artikels -> 80 % (abhängig von der Gestaltung des Seitenlayouts)
- Layouterkennung -> 87 % (abhängig vom Zustand/Qualität der einzelnen Elemente der Seite)
- OCR Erkennung -> 99,8 % (abhängig von der Druckqualität)

Wir produzieren die Einzelartikel auf Wunsch digiPaper-konform, msh Stuttgart, oder an DC-X angepasst.

www.prepress-systeme.de