

Vorgehen beim Scannen/Erfassen von Dokumenten für ein digitales Archiv

Von: Dr. Lothar Seveke, HTG

Wie immer, wenn man eine größere Arbeitsmenge vor sich hat, lohnt es sich, mehr Aufwand in die Vorbereitung und die Erprobung des besten Vorgehens zu stecken, um hinterher Zeit zu sparen.

Scanner oder Kamera?

Alle Erfassungsmethoden haben ihre Vor- und Nachteile, hochmoderne Automatikscanner kommen für uns sicherlich nicht infrage. Uns bleibt eigentlich nur übrig, entweder mit dem Flachbettscanner, A4 oder A3, oder einer Digital-Kamera zu arbeiten.

Normale Flachbettscanner sind nur für Einzelblätter oder dünne Hefte bzw. Zeitschriften geeignet. Für dickere Bücher bzw. gebundene Zeitschriften bräuchte man einen sogenannten Buchscanner, bei dem die gescannte Fläche bis an eine Längskante geht.

So etwas hat aber kaum jemand zu Hause, genauso wenig wie einen A3-Scanner, mit dem man auch bei A4-Dokumenten eine Doppelseite auf einmal scannen kann, was natürlich effektiver ist.

Mit einem Scanner dauert die eigentliche Erfassung länger als mit einer Kamera, dafür ist die Aufnahme immer scharf, und man braucht keine zusätzliche Beleuchtung. Spiegelungen auf der Oberfläche kommen normalerweise auch nicht vor.

Wenn man sich den Ablauf genau überlegt, kann man während der Laufzeit des Scanners schon das vorhergehende Bild bearbeiten und geordnet abspeichern, sodass man auch schnell arbeitet.

Wer hat und treibermäßig noch kann, sollte einen der älteren "dicken" Scanner (zum Beispiel aus der Scanjet-Serie von HP) verwenden und nicht die ultraflachen (zum Beispiel aus der Lide-Serie von Canon). Die dickeren schaffen aufgrund des dort verwendeten Aufnahmesystems eine wesentlich höhere Tiefenschärfe, sodass auch bei dickeren Büchern der Bruchteil in der Mitte der Doppelseite noch ausreichend scharf wird und man das Buch nicht zu sehr vergewaltigen muss.

Egal womit man arbeitet, sollte man standardmäßig Graustufen wählen und Farbe nur dort, wo es unbedingt nötig ist. Beim Scanner hieße die Einstellung also: Graustufen mit 300 dpi bzw. Farbe mit 300 dpi (nicht schwarz/weiß wählen), an der Kamera wählt man schwarz-weiß- bzw. Farbaufnahme. Mit der Auflösung ist es bei der Kamera etwas komplizierter. Für eine erfolgreiche OCR (optical character recognition) sind hier auch 300 dpi erstrebenswert. Bei dem gängigen Aufnahmeformat A4 (Doppelseite A5) würde das etwa einer Auflösung von 3500 x 2480 entsprechen. Man sollte also auf jeden Fall ein Aufnahmeformat 4:3 und dort zum Beispiel 12 MPixel (3984 x 2988) oder etwas in der Nähe wählen.

Ich arbeite nicht mit einer Digitalkamera sondern mit einem Smartphone bzw. dem iPad. Dort gibt es Anwendungsprogramme zum Einscannen von Dokumenten, die allerdings meist, ohne dass man es beeinflussen kann, nur mit einer Auflösung von 200 dpi arbeiten.

Das ist optimal geeignet, um mal schnell freihändig ein paar Seiten zu erfassen. Wenn es wie hier aber darum geht, z.B. Zeitschriften jahrgangsweise einzuscannen, und man nicht mit dem Scanner arbeiten kann oder will (gebundene Zeitschriften oder zu langsam), sollte man sich ein einfaches Stativ für die Digitalkamera oder das Smartphone basteln. Ich nehme dafür einen verstellbaren Gelenkarm mit Tischklemme aus der Beleuchtungstechnik, der vorne eine Platte als Auflage für das iPad hat. Auf der Tischplatte darunter kennzeichnet man sich den Rahmen, in dem die Zeitschrift liegen sollte. Die meisten der digitalen Aufnahmegeräte haben Autofokus, sodass man nach passender Höheneinstellung in schneller Folge umblättern und neu auslösen kann.

Weitere Randbedingungen

Die zu erfassenden Seiten sollten möglichst plan liegen, nicht unbedingt wegen der Schärfe, da eine Aufbeulung meist noch im Bereich der Tiefenschärfe liegt, sondern wegen der verzerrungsfreien Darstellung der Textzeilen, die für ein gutes OCR-Ergebnis wichtig ist.

Bei dünnen Heften ist dies meist kein Problem, bei dickeren und Büchern schon.

Man kann eine Glasscheibe auflegen und anpressen, die aber entspiegelt sein sollte, damit man sich nicht selbst auf jedem Bild sieht bzw. Reflexionen von der Beleuchtung erscheinen. Entspiegelte Glasscheiben gibt es für Bilderrahmen oder auch sonst beim Glaser. Schlechte Qualitäten von Ent-

spiegelungen erzeugen unter Umständen einen Grauschleier, was man auch beachten sollte. Um schneller arbeiten zu können, macht es sich gut, die Glasscheibe über ein gelenkiges Scharnier, das sich an die Dicke der Vorlage anpassen kann, wegklappen oder besser wegschieben zu können. Die homogene Ausleuchtung der Vorlage ist von großer Wichtigkeit für ein gutes Aufnahmeergebnis. Ich arbeite am liebsten mit diffusem hellem Tageslicht, was man leider nicht immer und überall zur Verfügung hat.

Mit Blitzlicht habe ich keine guten Erfahrungen. Zur Not tut es die Raumbeleuchtung oder ein bis zwei Schreibtischlampen von der Seite. Digitalkameras und Smartphones kommen scheinbar mit relativ wenig Licht aus. Man sollte jedoch beachten, dass die Beleuchtungsregelung zu Lasten der "Feinkörnigkeit" des Bildes geht und deshalb lieber die Automatik abschalten, wenn das möglich ist. Man kann dann auch viel besser erkennen, ob wirklich genügend Licht für eine helle, kontraststarke Aufnahme vorhanden ist.

Nachbearbeitung

Mit dem Flachbettscanner (oder der einfachen Digitalkamera) wird man meist das Rohergebnis gleich speichern und hinterher separat bearbeiten, wozu man, wie oben gesagt, die Pausen durch den Scannerlauf ganz gut nutzen kann.

Die Softwarelösungen in Smartphones und Touchpads bieten aber oft interessante Werkzeuge, um das Bild schon vor dem Abspeichern zum Beispiel zu entzerren und zu beschneiden. Günstig ist auch, wenn das Programm die Bilder gleich automatisch in einer Cloud oder auf dem PC speichert, wo sie bearbeitet werden können.

Die Bilder sind beim Abspeichern in geeigneter Weise zu benennen. Bitte verwendet dafür nur kleine Buchstaben ohne Umlaute, Ziffern und als einzige Sonderzeichen minus - und Unterstrich _. Wenn man also gerade den Jahrgang 1970 der Zeitschrift Poseidon erfasst, wäre folgende Benennung für die Seite 46 und 47 aus dem Heft 2 günstig:

pos-1970-02-46.jpg

Als erstes Ordnungskriterium nach dem Namenskennzeichen nehmen wir immer die Jahreszahl. Führende Nullen bitte immer so einsetzen, dass eine konstante Stellenzahl für die entsprechende Position entsteht, also nicht:

pos-1970-2-46.jpg

da sonst das Heft 10 vor dem Heft 2 in der Sortierreihenfolge erscheint.

Die einzelnen Jahrgänge kann man in Unterverzeichnissen zusammenfassen. Trotzdem sollte auch der Jahrgang im Namen enthalten sein, für den Fall, dass einmal eine Datei aus dem Verzeichnis entnommen wird.

Die Bilder können als JPG gespeichert werden, wobei die Komprimierung auf nicht geringer als 90 % eingestellt sein sollte.

Aufnehmen sollte man die Vorlage mit genügend breitem Rand, etwa 1 cm, damit nicht versehentlich Seiteninhalte abgeschnitten werden. Der Rand sollte aber auch nicht zu breit sein, damit der Verlust an Auflösung nicht zu groß ist.

Erster Schritt der Nachbearbeitung ist dann also das randlose Beschneiden, unter Umständen verbunden mit einer Entzerrung oder Ausrichtung der Lage.

Die so erhaltenen Bilder, jedes etwa zwischen 1 und 2 MByte groß, könnten so schon an das Archiv übergeben werden, wenn Ihr die noch genannten Möglichkeiten der Weiterverarbeitung nicht habt.

Um das endgültige Format für das Archiv zu erzeugen, sind noch zwei wesentliche Arbeitsschritte erforderlich. Wir wollen jedes Heft einer Zeitschrift als eine PDF-Datei abspeichern. Man muss also aus den Bildern der Doppelseiten des Heftes eine PDF erzeugen.

Dies geht sehr günstig zum Beispiel mit dem kostenlosen Programm IrfanView.

Damit erhält man entsprechend dem bisherigen Vorgehen eine bildbasierte PDF im Querformat in der Größe 25-50 MByte.

Auch dieses Format könntet ihr an das Archiv schicken, wir nehmen dann mit unseren Möglichkeiten die unten genannten nächsten Schritte noch vor.

50 Mbyte je Heft sind relativ viel Speicherplatz, den man aber noch reduzieren kann, ohne wesentlich an Qualität zu verlieren. Außerdem besteht diese PDF ja aus Bildern, man kann in ihr also nicht nach Stichwörtern suchen, die einen interessieren.

Mit der Software Adobe professional ist es möglich, in so einer Datei eine optische Zeichenerkennung (OCR) durchzuführen. Dadurch wird das bisherige Bild des Textes in wirklichen Text umgewandelt, in dem man suchen kann.

Außerdem verringert sich dadurch der Speicherbedarf für eine Heft-PDF auf etwa 10-15 MByte.

So eine textbasierte PDF, von Euch geliefert oder von den Archivaren aus Euren Scans oder Fotos erzeugt, wird dann im Archiv gespeichert.

Eine textbasierte PDF erkennt man übrigens daran, dass man darin mit der Maus einzelne Wörter markieren und gegebenenfalls herauskopieren kann.

Eine OCR mit den Einzelbildern vor der Zusammenfassung zur PDF durchzuführen, zum Beispiel mit einem OCR-Programm, wie es oft zum Scanner mitgeliefert wird, ist ein müßiges Unterfangen. Der Arbeitsaufwand ist viel zu hoch und das Ergebnis meist doch nicht befriedigend, da vor allem das Layout nur sehr schlecht dem Original entspricht

Die Durchsuchbarkeit der PDF, also die Auffindbarkeit gesuchter Wörter, ist von dem Ergebnis der OCR abhängig und dieses wiederum von der Qualität der Vorlage und der der Erfassung. Man sollte also in so einer aus gescannten Bildern erzeugten PDF nicht erwarten, dass jedes Wort gefunden wird. Aber schon eine große Wahrscheinlichkeit dafür ist eine Hilfe bei der Suche.

Alle Daten, die Ihr an das Archiv liefern möchtet, seien es die Bilder der Seiten oder eine grafikbasierte PDF eines Heftes oder gar die textbasierte PDF, könnt ihr selbst hochladen.

Man ruft folgende Adresse im Browser auf:

www.wetransfer.com

gibt die Empfängeradresse archiv@historische-tauchergesellschaft.de und seine eigen Email-Adresse an und kennzeichnet auf seinem PC die entsprechenden Dokumente.

Ein Archivar kann sich die Dokumente dann von weTransfer herunterladen.

Damit Euer Fleiß auch gewürdigt werden kann, fasst bitte die Dateien in einem ZIP-Archiv Eures Namens zusammen.

Das sind so meine Vorstellungen und Erfahrungen.

Wenn Ihr andere habt oder sonst bessere Vorschläge, lasst es uns, die Archivare, wissen.