Nicolas Flores-Herr
Monika Pfitzner
Jörg Räuber

Chancen von Verfahren zur Massendigitalisierung von Büchern

Bericht über den Test eines Buchscanroboters; Perspektive von Hochdurchsatzdigitalisierung zur Bestandserhaltung

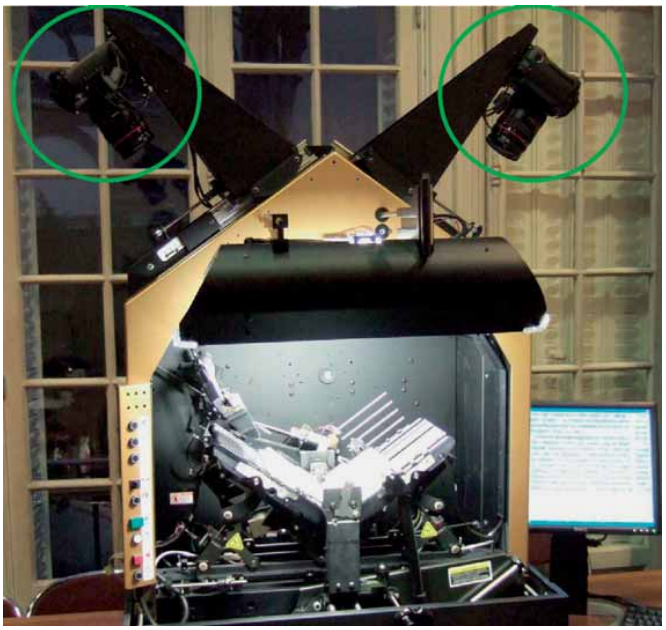
Seit geraumer Zeit wurden in der Deutschen Nationalbibliothek (DNB) in Leipzig Überlegungen zur Digitalisierung unter dem Aspekt der Bestandserhaltung angestellt. Im Leipziger Bestand befinden sich mehr als 3.200 Objekte, deren Benutzung aufgrund des Erhaltungszustandes nicht mehr möglich ist. Eine restauratorische Bearbeitung ist bei diesen Exemplaren nur über das Papierspaltverfahren denkbar. Da die finanziellen Mittel für die Bestandserhaltung begrenzt sind, kann diese Erhaltungsmaßnahme jedoch derzeit nicht angewendet werden. Eine weitere kritische Bestandsgruppe stellen die im Spirit-Umdruckverfahren hergestellten DDR-Dissertationen[1)] dar, die durch Pigmentverlust bedroht sind. Durch Digitalisierung wäre eine dauerhafte Sicherung und Zugänglichkeit der Inhalte gewährleistet. Außerdem würden sich durch die elektronische Erfassung der Inhalte weitere Recherchemöglichkeiten für den Nutzer ergeben (z. B. Volltextsuche, Bereitstellung von Inhaltsverzeichnissen).

Perspektive von Hochdurchsatzdigitalisierung für die European Digital Library (EDL)

Im Kontext der EDL sollen in jedem Partnerland Kompetenzzentren für Digitalisierung entstehen. Ein Fokus des Projekts wird auf Massenverfahren zur Digitalisierung liegen. Daher werden von vielen europäischen Nationalbibliotheken z. B. der Bibliothèque nationale de France (BnF) sowie deutschen Digitalisierungszentren Studien und praktische Tests zum Thema Hochdurchsatzdigitalisierung geplant und durchgeführt.

Weitere Beweggründe als erste kulturelle Einrichtung Deutschlands einen Hochdurchsatzscanner zu testen, bestanden im Gewinn von Erkenntnissen über die Handhabung der Bücher durch den Scanner, der Messung des effektiven Durchsatzes (digitalisierte Seiten pro Tag) sowie die Erforschung der Qualität und die massenhafte Verarbeitung der Scans. Da die Firma BancTec im Jahr 2006 als einziger Dienstleister Deutschlands über einen Buchscanroboter verfügte, welcher vollautomatisiert Buchseiten umblättert (APT 2400 der Firma Kirtas), wurde die Firma beauftragt die Teststellung durchzuführen.

Teststellung vollautomatischer Buchscanroboter



Der APT 2400 digitalisiert das eingelegte Buch durch zwei digitale Spiegelreflexkameras (Canon EOS-1Ds Mark II). Die SLRs sind mit einem Firewireanschluss an je einen PC angeschlossen. Die Dateien werden auf einem »shared« Laufwerk zentral gespeichert. Nach dem Umblättern der Seite werden die Kameras simultan ausgelöst. Die Auslösung der Kameras erfolgt entweder automatisiert oder manuell per Bedienfeld.

Es wurde vereinbart, dass im Rahmen eines Projektes 250.000 Seiten aus den Beständen

der DNB in Leipzig digitalisiert werden sollen. Die Auswahl der zu digitalisierenden Bestände erfolgte sowohl unter dem Benutzungsaspekt als auch zur Untersuchung, welche Buchvorlagen für eine Verarbeitung durch den Kirtas Scanner geeignet sind. Berücksichtigt wurden alle Bereiche des Hauses, sowohl der allgemeine Bestand als auch Objekte aus dem Deutschen Buch- und Schriftmuseum, aus den Sonder- und Spezialsammlungen und aus dem Hausarchiv.

Eine Überlegung war, stark frequentierte Bestände nicht mehr im Original für die Benutzung zur Verfügung zu stellen, sondern künftig das Digitalisat in den Lesesälen zu nutzen. Durch die Schonung des Originals wäre damit auch eine längere Haltbarkeit der Papierausgabe gewährleistet.

Ein weiterer Ansatz ging davon aus, an möglichst vielen Dokumenttypen die Möglichkeiten und Grenzen der Digitalisierung und Texterkennung (Optical Character Recognition, OCR) zu testen.

Für die Teststellung wurden verschiedene Vorlagengrößen, Papierqualitäten, Papiergewichte, schwarz-weiße und farbige Vorlagen, brüchige Papiere etc. in einer Vorauswahl zusammengestellt. Insgesamt wurden Bücher mit rund 294.131 Seiten bereitgestellt.

Ausschlusskriterien

Bekannt waren im Vorfeld technisch bedingte Ausschlusskriterien für Veröffentlichungen mit folgenden Eigenschaften:

- Format

kleiner als 11,5 cm x 18 cm, größer als 28 cm x 35,5 cm, Buchrücken dicker als 10 cm.

- Papiergewicht

leichter als 49 g/m², schwerer als 120 g/ m².

- Zustand

lose Seiten, lose Beilagen, Seiten mit großen Löchern, brüchiges Papier, fehlendes Buchcover, Klappendeckel hält die Buchseiten nicht mehr.

- Sonstiges

Objekte mit eingelegten Karten, Objekte mit Seiten zum Auffalten.

Bei der Bereitstellung der Testbände und der Einrichtung des entsprechenden Geschäftsganges stellten sich weitere Einschränkungen heraus. So konnten z. B. fortlaufende Sammelwerke sowie Bände, die mehrere bibliografisch selbstständige Werke enthalten (Container) nicht mit einer Akzessionsnummer versehen werden, da bisher keine eindeutige Verknüpfung mit einem Datensatz herstellbar ist.

Durchführung der Teststellung

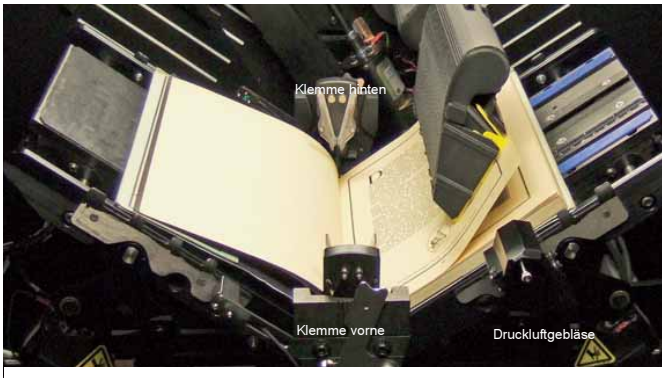
Seitens des Herstellers Kirtas wurde die Bruttoleistung des APT 2400 für einen achtstündigen Arbeitstag auf 19.200 Seiten beziffert. Der realistische Wert lag nach Aussage der Firma BancTec bei 70 %; mithilfe des Geräts sollten demnach rund 13.000 Buchseiten pro Tag digitalisiert werden können.

Die Digitalisierung wurde in den Räumen der DNB in Leipzig durchgeführt. Das Bedienpersonal (eine Person) für den APT 2400 wurde von der Firma BancTec zur Verfügung gestellt. Eine Vorauswahl wurde sowohl im Hinblick auf die Digitalisierung möglichst verschiedenartiger Werke in unterschiedlichem Erhaltungszustand als auch unter der Bedingung mehrerer Testtage zur Messung des maximalen Durchsatzes vorgenommen.

Somit wurde einerseits die Grundlage geschaffen, die für zukünftige Bestands- bzw. Inhaltssicherung notwendigen physikalischen Parameter der Bücher abzustecken und andererseits den Nettodurchsatz der Maschine zu messen.

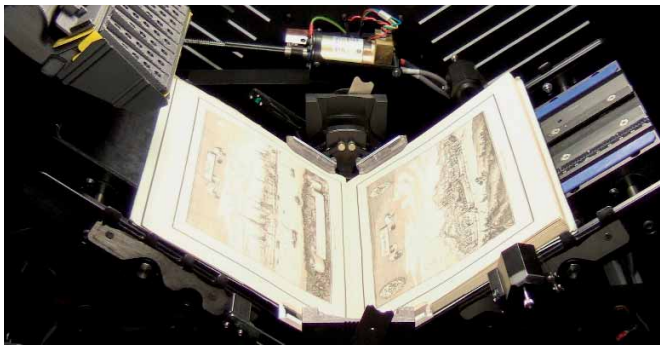
Funktionsweise des Gerätes

Der Kirtas APT 2400 blättert die Seiten mithilfe eines Roboterarms um, welcher durch einen Unterdruck die Buchseiten ansaugt. Der Unterdruck ist einstellbar und an die Vorlage anpassbar. Durch optimale Wahl der Parameter kann das gleichzeitige Umblättern von zwei Buchseiten verhindert werden. Dies erfordert allerdings Erfahrung und Fingerspitzengefühl des Operators.



Umblättervorgang: Eine Buchseite wird durch den Roboterarm angesaugt.

Im direkten Anschluss an den Umblättervorgang wird der Roboterarm weggefahren und transparente Klemmen aus Kunststoff nahe der Buchmitte drücken die Seiten nach unten. Dieser Druck ist ebenfalls einstellbar. Sobald die Klemmen die Buchseiten fixiert haben, werden die Buchseiten digitalisiert.



Digitalisierungsstellung: Durch das Herunterfahren der Klemmen werden die Buchseiten heruntergedrückt, um die geometrischen Verzerrungen möglichst gering zu halten. Der Druck der Klemmen kann variiert werden.

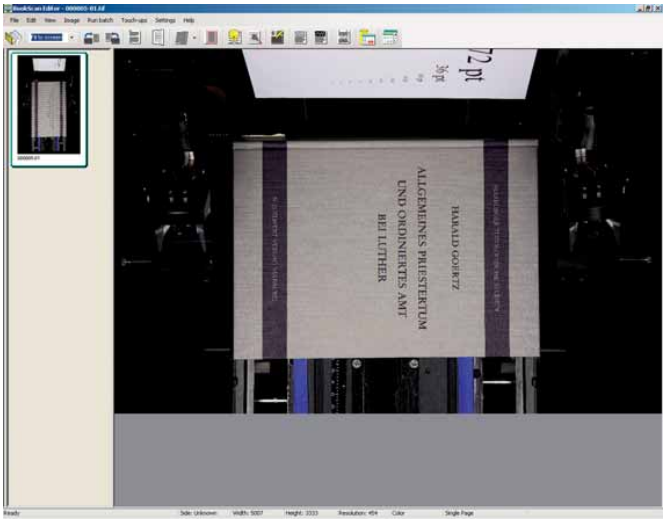
Rüstvorgang

Bevor der Digitalisierungsvorgang beginnt, muss die Maschine für jedes Buch angepasst werden.

Der Digitalisierungsvorgang wird in der Regel vollautomatisch durch die Kirtas Software eines angeschlossenen PCs gesteuert. Allerdings kann das Gerät nicht detektieren, wenn Seiten doppelt umgeblättert werden oder wenn keine Seite umgeblättert wird. Aus diesem Grund muss der Operator während der Digitalisierung andauernd den Umblättervorgang beobachten. Diese Tätigkeit ist sehr ermüdend und nicht über einen langen Zeitraum möglich.

Scantechnologie

Der Kirtas APT 2400 digitalisiert Buchseiten mithilfe zweier Canon EOS-1Ds Mark II Digitalkameras (16,7 Megapixel), welche nach jedem Umblättervorgang je eine Buchseite fotografieren. Die Kameras erzeugen zwei Typen digitaler Bilder: Raw-Bilder, so genannte »Digitale Negative« und JPEG-Bilder »Digitale Positive«.



Ein Screenshot der BookScan Editor Pro™ Software: Im oberen Bereich sind die Bedienelemente für die Bildbearbeitung zu sehen. Für jedes Buch müssen je Kamera die Bildbearbeitungsschritte durchgeführt werden. Die Einstellungen der Bildbearbeitungsschritte werden dann für jede Buchseite übernommen und im Batchverfahren angewendet.

Die Bildoptimierung erfolgte durch die BookScan Editor ProTM Software (Kirtas Technologies; Victor, New York) und umfasste folgende Arbeitsschritte:

- Drehen der Images,
- Entfernen der Klemmen,
- Entfernung der Objektivverzeichnung,
- Entzerren des Bildes,
- Einstellung von Helligkeit/Kontrast,
- Schärfen des Bildes,
- Entfernen des Buchseitenhintergrundes und Nachschwärzen des Textes,
- Automatisches Erkennen der Bilder.

Die Erzeugung von elektronischem Text aus Scans der Buchseiten

Einige der digitalisierten Bücher wurden mit der gelieferten Texterfassungs/OCR-Software (»OCR Manager« von Kirtas Technologies) bearbeitet und die entsprechenden Ergebnisse als PDF und als XML-Datei zur Verfügung gestellt.

Zu einem großen Teil konnte Bildbearbeitung und Texterfassung nicht während der Projektlaufzeit realisiert werden, da sowohl Kirtas BSE Pro schlechte Ergebnisse lieferte, als auch »OCR Manager« fehlerhaft war, was zu enormen Zeitverzögerungen führte.

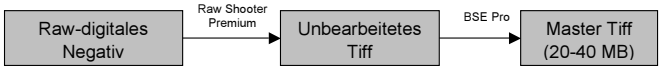
Betriebsmodi

Im Rahmen der Teststellung wurden insgesamt zwei Betriebsmodi des Scanners untersucht:

DNB-Raw-Workflow – Hohe Bildqualität durch Verwendung Digitaler Negative

Dieser Workflow war nicht durch die Firma Kirtas vorgesehen, aber aus zwei Gründen wurde es als notwendig erachtet, mit Raw-Formaten zu arbeiten:

- Deutlich bessere Bildqualität und wirklichkeitsgetreuere Helligkeits- und Farbwiedergabe,
- Aktuelle Diskussionen über die Langzeitarchivierung von digitalen Negativen (Adobe DNG, Open Raw).



Der Raw-Workflow wurde durch die DNB entwickelt. Durch die Verarbeitung/»Entwicklung« eines digitalen Negativs außerhalb der Kamera mittels der PC-Software »Raw Shooter Premium« konnte eine erhebliche Qualitätsverbesserung sowohl bei der Wiedergabe von Helligkeitsstufen als auch bei den Farben des Originals erzielt werden. Durch die Bearbeitung, Speicherung und Übertragung großer Dateien liegt die Bearbeitungsdauer mit den bei der Teststellung genutzten PCs und Netzwerken bei über einer Minute pro Bild.

Kirtas-JPEG-Workflow – Höchster Durchsatz durch die Verwendung Digitaler Abzüge / Positive

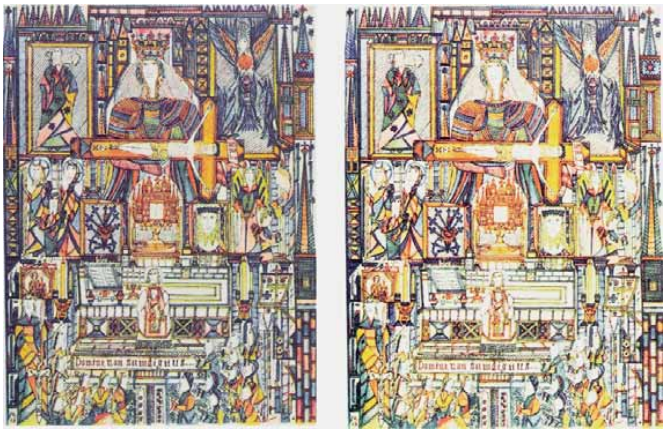
Dieser Workflow entspricht den Vorgaben des Herstellers Kirtas; er ist aus zweierlei Gesichtspunkten heraus nicht für den kulturellen Bereich interessant:

- Das JPEG-Kompressionsverfahren ist verlustbehaftet; Bildinformation geht verloren.
- Die Helligkeits- und Farbwiedergabe der JPEG Bilder ist durch interne Parameter der Kamera festgelegt. Korrekturen an den JPEG Bildern sind im Gegensatz zur Verwendung von Raw-Formaten immer verlustbehaftet.

Kirtas-JPEG-Workflow: Hier werden von der Kamera verlustbehaftete JPEG-komprimierte Bilder ausgegeben. Die Qualität der Master-Tiffs ist deutlich schlechter als im DNB-Raw-Workflow.

Qualitätsvergleich

Der DNB-Raw-Workflow unterscheidet sich vom Kirtas Workflow durch eine verbesserte Wiedergabe von Farben und Helligkeitsstufen der Vorlage.



Das Bild auf der linken Seite ist deutlich dunkler als das Bild auf der rechten Seite und besitzt weniger kräftige Farben. Allerdings ist das Bild auf der linken Seite wesentlich farbtreuer; dies ist allerdings nur zu beobachten, wenn die Papiervorlage unter Standardlichtbedingungen mit dem Bild eines kalibrierten Monitors verglichen wird. Weiterhin gehen in der rechten Abbildung Zeichnungsdetails in großem Maße verloren.

Der einzige Nachteil des DNB-Raw-Workflows lag in der langsamen Nachbearbeitung der Scans und Übertragung der großen Dateien (50 MB/Datei) über das Netzwerk. Der Durchsatz lag selbst bei Stapelverarbeitung über Nacht bei unter 1.500 Bildern/Tag.

Ergebnisse

Es konnten nicht alle exemplarbezogenen Fragen im Rahmen des Tests beantwortet werden. So wurden Dissertationen, welche im »Ormig-Verfahren« hergestellt wurden, nicht einbezogen, da das Druckbild zum Teil nicht einmal mit dem bloßen Auge zu erkennen war. Eventuell könnten die Dissertationen in Zukunft unter Infrarotbeleuchtung mit speziell angepassten Kameras digitalisiert werden.

Die Digitalisierung mehrbändiger Werke war nicht möglich, weil sich für ein physikalisches Objekt (z. B. Buch) innerhalb eines Bandes kein Identifikator finden ließ, der einen eindeutigen Zusammenhang zwischen Katalogeintrag und physikalischem Objekt herstellte.

Nach Abschluss des Projektes waren insgesamt 417 Bücher (mit 201.460 Seiten laut Katalog) digitalisiert. Tatsächlich entstanden 195.268 digitalisierte Seiten, davon 15.446 Seiten in Frakturschrift.

Digitalisierte Materialien

Sofern die Digitalisierung mithilfe des Kirtas-JPEG-Workflows durchgeführt wurde, konnte ein Durchsatz von 10.000 Seiten/Tag erreicht werden (Spitzenwert: 12.785 Seiten/Tag). Zur Steigerung des Tagesdurchsatzes war es von großem Vorteil, wenn Bücher ähnlichen Formats und mit großer Anzahl von Seiten digitalisiert wurden. Auf diese Weise konnten die in der Regel zeitintensiven Rüstphasen auf ein Minimum beschränkt werden.

Durchsatz

Die Bildbearbeitungs- und Texterfassungssoftware der Firma Kirtas wies erhebliche Mängel in den Bereichen Robustheit, Useability und Bearbeitungsqualität auf, die für beträchtliche Zeitverzögerungen im Projektablauf sorgten. Solange diese Mängel nicht durch Kirtas Technologies behoben werden, sollten für zukünftige Dienstleistungen zwei Operatoren an einer APT 2400 arbeiten. Um

einen Durchsatz von 10.000 Seiten pro Tag zu gewährleisten, muss nach dem heutigen Stand der Technik ein Operator die Bildbearbeitung/OCR-Erfassung und ein Operator den APT 2400 steuern.

Kirtas APT 2400 ist zu verbessern

Es ist festzuhalten, dass ebenfalls der APT 2400 verbessert werden muss, um den Durchsatz zu steigern. Zwar war der Wartungsbedarf gering, sodass es zu keinen nennenswerten Zeitverzögerungen während des Projekts kam, aber es war nicht möglich, während des Scanvorgangs die Maschine unbeobachtet zu lassen, da z. B. die Mitnahme von zwei Buchseiten durch den Roboterarm regelmäßig auftrat. Deshalb war es dem Operator nicht möglich, gleichzeitig Batchprozesse zur Bildbearbeitung und Texterfassung anzustoßen.

Keines der digitalisierten Bücher erlitt durch die Handhabung der Maschine einen nachweisbaren Schaden. Hier muss hinzugefügt werden, dass zwar ältere Bücher digitalisiert wurden, aber Werke, die nicht mehr ohne besondere Vorsichtsmaßnahmen angefasst werden konnten, nicht berücksichtigt wurden.

Fazit

Prinzipiell ist der APT 2400 geeignet, in zukünftigen Projekten durch Digitalisierung zur Bestandserhaltung beizutragen. Kritische bzw. extrem beschädigte Bestände wären noch gesondert zu testen.

Welche Technologien sind zur Weiterverarbeitung der Digitalisate notwendig?

Ausblick

Die Erfahrung mit der Kirtas Bildbearbeitungssoftware hat gezeigt, dass die Verarbeitung von 10.000 Bildern am Tag nur erfolgen kann, wenn Software zum Einsatz kommt, welche die grundlegenden Bildbearbeitungsschritte (Zuschneiden, Entzerren, Entfernen von Artefakten, Anpassung von Kontrast, Helligkeit und Gamma, etc.) automatisiert durchführt.

Anmerkungen

1
Spirit-Umdruckverfahren für Auflagenhöhen von 50 bis 100 Exemplaren, das vorwiegend in der ehemaligen DDR für die Vervielfältigung von Dissertationen verwendet wurde (Ormig).