Machbarkeitsuntersuchung: Speicherung digitaler Daten auf Mikrofilm durch Standardtechnologien

Steffen W. Schilke

Hessische Zentrale für Datenverarbeitung
E1 – DMS und zentrale Fachanwendungen
Mainzer Strasse 29
65185 Wiesbaden
S.Schilke@hzd.hessen.de

Abstract: Im Rahmen von Fachanwendungen und E-Government müssen digitale Daten einer Langzeitarchivierung unterzogen werden. Da digitale Daten normalerweise digital gespeichert werden, hat man neben der Problematik der langzeitarchivierungsfähigen Dateiformate auch das Problem der nachhaltigen Langzeitarchivierung auf Medien. Wenn diese auf digitalen Medien stattfindet, ist mit einer regelmäßigen Medienmigrationen zu rechnen. Die Machbarkeitsuntersuchung soll feststellen, ob die Ablage digitaler Daten auf Mikrofilm eine Alternative zur Speicherung auf digitalen Medien darstellen kann. Dabei sollen die Daten nicht in der Form von Images abgelegt werden, sondern in einem Format, welches auch wieder in ein digitales Format umgesetzt werden kann.

1 Einführung

Nicht nur im Rahmen der Aussonderung entstehen in Fachanwendungen und in e-Government Anwendungen Daten, die Langzeitarchiviert werden müssen. Diese Daten werden zurzeit in langzeitarchivierungsfähige Formate [BSIa] umgesetzt und auf digitalen Medien abgelegt. Die aktuellen Speichermedien sind z.B. Bänder (Tape), festplattenbasierende Speichersysteme (SAN, NAS, DAS, …), WORM-Medien oder spezielle Langzeitarchivierungsspeicher wie die EMC Centera oder entsprechende Systeme von z.B. NetApp, SUN (HoneyComb), etc..

Diese Speichermedien zeichnen sich durch eine begrenzte Lebensdauer bzw. Laufzeit aus und man muss bei der Planung von Archivierungssystemen eine Medienmigration bzw. Auffrischung geplant werden [BSIb, BSIc, BSId]. Dies ist bedingt z.B. durch den technologischen Lebenszyklus bzw. die technische Lebensdauer der Systeme. Im Rahmen dieser Untersuchung werden die ebenso notwendigen Punkte Formatmigration oder Singnaturerneuerung nicht betrachtet. Mikrofilm als Speichermedium hingegen hat durch seine Haltbarkeit von 100+ Jahren den Vorteil weniger häufig einer Medienmigration unterzogen werden zu müssen.

Der Punkt Formatmigration sollte durch die Verwendung ausgezeichneter Langzeitarchivierungsformate eine geringere Notwendigkeit bedingen, als die öfters vorkommende technisch notwendige Medienmigration.

2 Ansätze zur Speicherung auf Mikrofilm

Zur Ablage von digitalen Daten auf Mikrofilm sind verschiedene Verfahren denkbar. Im Rahmen des Projektes Arche [WS07] wurde ein Ausblick auf die digitale Speicherung von digitalen Daten gegeben. Diese Methode wird zurzeit im BMWI-Projekt Millennium im Projekte „Bits on Film“ untersucht [Bi06] und soll auf einer Spezial-Hardware realisiert werden. Dabei soll eine Ablage mittels eines speziellen Farbmikrofilmlaserschreibers erfolgen. Das bedingt dann auch eine entsprechende spezialisierten Hardware zum Rückdigitalisieren der abgelegten digitalen Daten.

Im Rahmen dieser ersten Machbarkeitsuntersuchung soll die Ablage von digitalen Daten auf Standard Mikrofilm mittels Standardtechnologien untersucht werden. Die im Rahmen der Tests eingesetzten Systeme bestehen aus dem Kodak Digital Archive Writer i9610 (DAW - Mikrofilmschreiber) und einem Kodak 3000 DSV (bzw. Minolta ES-7000 - Mikrofilmscanner) mit der jeweils dazugehörigen Software (PowerFilm bzw. AWIS/KIWI). Die Mikrofilme werden mit einer Standardentwicklungseinheit für Mikrofilme (Kodak Prostar) entwickelt. Die eingesetzten Mikrofilme sind 30,5 Meter lange 16mm Schwarz/Weiß-Mikrofilme, die auch für die analoge Archivierung von Bilddaten (Images) durch eingesetzt werden. Diese Technologie befindet sich schon seit vielen Jahren im erfolgreichen Einsatz im Langzeitarchivierungsumfeld (seit ca. 100 Jahren). Um die digitalen Daten auf einen Mikrofilm abzulegen, müssen die Daten in ein passendes Bildformat überführt werden. Da keine Bildrepräsentation/Abbild der Daten abgelegt werden soll, sondern die digitalen Daten selbst, muss das abzulegende Format wieder in digitale Informationen umgewandelt werden können.

2.1 Ablauf der Untersuchung

Die Daten, welche im Rahmen der Untersuchung genutzt werden, liegen z.B. in den Formaten Text (TXT und CSV), TIFF, PDF und PDF/A (ISO 19005-1:2005) vor. Für jede der Quelldateien werden Hash Werte zur Überprüfung der Ergebnisse der Rückdigitalisierung vom Mikrofilm erzeugt. Die Quelldateien werden mit den entsprechenden Konvertern in die zu testenden textbasierenden / barcode-artigen Formate (Bilder/Images) umgesetzt und direkt wieder zurückgewandelt, um diesen Prozess vorab zu testen. Danach werden die erzeugten Bilder zum Schreiben auf Mikrofilm vorbereitet. Dies bedingt eine Formatkonvertierung, um ein (Bild/Image) Format zu erhalten, welches vom DAW auf den Mikrofilm geschrieben werden kann. Als Format kommt zum Schreiben TIFF (Tagged Image File Format) zum Einsatz. Um eine breites Spektrum als Testdaten zu erhalten werden dabei verschiedene Varianten der Wandlung durch die untersuchten Programme genutzt.

Nach dem Schreiben des Mikrofilms wird dieser entwickelt und einer optischen Kontrolle unterzogen. Im Anschluss wird der Mikrofilm mittels des Mikrofilmscanners wieder rückdigitalisiert. Die gescannten Images werden den entsprechenden und benötigten Format(rückum)wandlungen unterzogen. Die erhaltenen Bilddaten werden dann wieder in ihre „digitale Urform“ zurückgewandelt und mittels der Hash Codes geprüft. Falls diese Prüfung positiv ausfällt, wird die erzeugte Datei mit der passenden Anwendung geöffnet und geprüft.

2.2 Textbasierende Formate

Die im Internet verwendeten Standardverfahren, um digitale Daten in Text umzusetzen, werden z.B. in Emails oder bei der Nutzung von Webservices eingesetzt. Im Rahmen des Tests werden die Format Base64, Binhex und z.B. UUencode untersucht. Die Testdateien werden mittels eines Open - Source Konverters in eine Textrepräsentation umgewandelt und über einen TIFF Druckertreiber (Microsoft Office Document Imaging Writer) als TIFF Dateien ausgegeben. Dabei werden verschiedene Auflösungen und Schriftgrößen einer Standardschrift genutzt, um die Auswirkungen untersuchen zu können.

2.3 Barcode-artige Formate

Barcode-artige Verfahren setzen die digitalen Daten in eine Art Ganzseiten-Barcode um. Dieser Barcode wird von den zwei untersuchten Open Source Verfahren Paperback und Optar hauptsächlich zur Speicherung von Daten auf Papier vorgeschlagen. Die erzeugte barcode-artige Form liefert eine Art 2D-Barcode in Schwarz/Weiß. Die abgelegten Daten sollen dann über einen Scanner vom Papier rückdigitalisiert werden und dann wieder in eine digitale Repräsentation umgewandelt werden können. Dabei wird von den Anbietern der Programme empfohlen, eine höhere Auflösung für das Einscannen der Barcode-artigen Seiten zu benutzen. Diese Verfahren werden in der vorliegenden Machbarkeitsuntersuchung zum ersten Mal auf ihre Ablagefähigkeit auf Mikrofilm untersucht.

Dabei ist zu beachten, dass bei der Ausbelichtung auf Mikrofilm ein Verkleinerungsfaktor genutzt wird, um die Seiten auf den Mikrofilm belichten zu können. Es muss auch beachtet werden, dass bei der Rückdigitalisierung vom Mikrofilm eine entsprechende Vergrößerung im Mikrofilmscanner vorgenommen wird. Dies hat natürlich Auswirkungen auf die Qualität der Bilder und das zu erzielende Ergebnis. Die vom Autoren vorgenommenen Anpassung der Ausgabe (z.B. Wandlung von der Optar Ausgabe PNG zu TIFF) bzw. des Paperback Ausgabeformates (Nutzung eines TIFF Druckertreibers) waren notwendig um die Bilddateien mittels des Mikrofilmschreibers auf Mikrofilm schreiben zu können. Das Programm Optar erzeugt eine Art 2D-Barcode und beugt Fehlern bei der Rückdigitalisierung durch die Verwendung eines Golay-Codes [Go49] vor. Der Golay-Code dient dazu, bis zu 3 Bit Fehler in den Daten korrigieren und 4 Bit Fehler erkennen zu können. Dazu werden 12 Datenbits innerhalb von 24 Bit kodiert. Das benötigte Dekodier-Programm von Optar (UnOptar) ließ sich nicht auf den zur Verfügung stehenden Windows Systemen kompilieren.

Das Programm PaperBack gibt die erzeugten Bilddateien über einen Druckertreiber aus und bietet auch eine Möglichkeit, die Bilddaten direkt als BMP Datei abzuspeichern. Der oben erwähnte TIFF-Druckertreiber wurde für den Test als Ausgabegerät verwendet. Dabei wurden die DPI –Werte (Dots Per Inch) 100, 200 und 300 DPI verwendet. Um Fehler bei der Dekodierung ausgleichen zu können, verwendet PaperBack eine Reed-Solomon Error Correction [RS60]. Zusätzlich besteht die Möglichkeit, eine Redundanz für die abgelegten Daten zu bestimmen. Diese wird im Rahmen der Untersuchung auf verschiedene Werte gesetzt (1:5 und 1:7). Dabei stellt ein Faktor von 1:5 eine Redundanz für 5 Datenblöcke dar. Das bedeutet, dass wenn auf 5 Datenblöcken einer nicht lesbar ist, kann dieser wiederhergestellt werden. Auch der Abstand zwischen den Punkten (dot size, Whitespace) wurde testweise auf verschiedene Werte eingestellt. Zusätzlich wurde die eingebaute Datenkompression von Paperpack verwendet.

2.4 Schreiben auf Mikrofilm

Im Rahmen der Untersuchung wurden die Bilddaten, welche die Bildrepräsentation einer digitalen Datei darstellen, als A3 quer bzw. Tabloid quer (11“ x 17“) erzeugt. Diese TIFF Dateien wurden dann mit der, von der Mikrofilmschreibersoftware empfohlener, Verkleinerung auf den Film geschrieben. Zum Schreiben wurde das Querformat ausgewählt, um beim Simplexschreiben des Mikrofilmes die zur Verfügung stehende Fläche des Mikrofilms optimal ausnutzen zu können. Dieses Verfahren wurde bei den textbasierenden Formaten und bei den beiden barcode-artigen Formaten verwendet. Abhängig vom vorliegenden Format (A3 quer bzw. Tabloid quer) lagen die Verkleinerungsfaktoren zwischen 40- und 50-facher Verkleinerung. Die vorliegenden Bilddateien wurden dann sequenziell auf den Mikrofilm geschrieben. Dabei verwendet der Kodak Digital Archive Writer eine LED-Zeile, die es ermöglicht, 3888 Punkte pro Zeile auf den Mikrofilm zu schreiben. Beim Erstellen der Bilddaten wurde darauf geachtet diesen Wert nicht zu überschreiten. Der Mikrofilmschreiber schreibt nur Schwarz/Weiß Bilder auf den Mikrofilm. Die zu schreibenden Bilder werden durch die Mikrofilmschreiber Software invertiert.

2.5 Rückdigitalisieren der Bilddateien vom Mikrofilm

Zur Rückdigitalisierung der Bilddateien wurde ein Kodak 3000 DSV Mikrofilmscanner verwendet. Dieser wurde auf die einzelnen Bilddateien eingerichtet; die Vergrößerung vom Film zusätzlich manuell auf die bestmögliche Vergrößerung für die Rückdigitalisierung eingestellt. Da diese Einstellung rein mechanisch über ein Drehrad funktioniert, ist es leider nicht möglich, die gewählte Vergrößerung auszulesen. Die verwendete Powerfilm Software steuert den Mikrofilmscanner an und liest die Bilddaten über die Optik und die Scann-Einrichtung ein und legt die Bilder als TIFF ab. Die verwendeten Einstellungen waren: Bild Invertieren (um das Originalbild wieder zu erhalten), 600 DPI und Graustufen. Hierbei wurde der Empfehlung der Anbieter der Systeme Optar und Paperback gefolgt. Die Bilder wurden dann abgelegt und in die benötigten Formate zur Auswertung durch die erzeugenden Systeme gewandelt (Paperback: BMP, Optar: PNG).

3 Auswertung der Studie

Diese erste Machbarkeitsstudie zeigte, dass es prinzipiell möglich ist, digitale Daten auf Mikrofilm abzulegen und als digitale Daten wiederherzustellen. Dabei muss auf die physikalisch bedingte Unzulänglichkeit der verwendeten Systeme Rücksicht genommen werden. Der Versuch, Daten in einem textbasierenden Format wie Base64, Binhex und UUencode abzulegen, scheiterte an der Qualität der OCR-Ergebnisse der rückdigitalisierten Daten. Da verschiedene Auflösungen (DPI des Images) und Schriftgrößen (z.B. Arial 22 Punkt) verwendet wurden konnte im Rahmen der Versuche festgestellt werden, dass die Ergebnisse besser werden, wenn die verwendete Schrift größer war und die Zeichenmuster für die verwendete OCR Engine einfach zu unterscheiden waren. Je besser, d.h. je klarer und „größer“ die erzeugten Images auch vom menschlichen Auge zu lesen waren, desto besser waren die Ergebnisse auch von der OCR zu interpretieren. Dabei waren Fehlermengen von 10+ Fehlern pro Zeile bei niedrigen Auflösungen mit „kleinen“ Schriften keine Seltenheit. Gut „lesbare“ Ergebnisse (hohe Auflösung und „große“ Schriften) lagen bei 1-2 Fehlern pro Zeile Text. Dies ist aber, durch die mangelnde Redundanz und Fehlerkorrektur bedingt, kein Ergebnis, welches diese Verfahren für eine Langzeitarchivierung von digitalen Daten empfehlen kann.

Die Auswertung von Barcode-artigen Formaten konnte im Rahmen dieser Machbarkeitsstudie nur mit Paperback durchgeführt werden, da UnOptar sich nicht auf einem Windowssystem kompilieren lies. Optar soll in einer weiteren Untersuchung betrachtet werden und mit den vorliegenden Ergebnissen verglichen werden. Abhängig von den bei Paperback verwendeten Einstellungen (Redundanz und Auflösung in DPI) konnten, dank der verwendeten Fehlerkorrektur und der verwendeten Redundanz, die abgelegten Daten in einigen Fällen zu 100% wiederhergestellt werden. Als Beispiel ist hier die Datei book.tif und ihre Rückwandlungen aufgeführt:

File : book.tif (Original) SHA-256 : 477CFAB5 3DC3EBF2 47D14E6C 2E8EC527 8E3C2649 CC664E33 0DBC0ED0 72CEA113
File : book_1.tif (Paperback 13 Seiten; 1:5 Redundanz; 200 DPI TIFF; 70% Dot Size) SHA-256 : 477CFAB5 3DC3EBF2 47D14E6C 2E8EC527 8E3C2649 CC664E33 0DBC0ED0 72CEA113 – 2293 good Blocks; 1100 Byte EEC Korrekturen
File : book_2.tif (Paperback 12 Seiten; 1:7 Redundanz; 200 DPI TIFF; 70% Dot Size) SHA-256 : 477CFAB5 3DC3EBF2 47D14E6C 2E8EC527 8E3C2649 CC664E33 0DBC0ED0 72CEA113 – 2209 good Blocks; 720 Byte EEC Korrekturen

Die Rückdigitalisierung funktionierte nur einwandfrei, wenn eine niedrige DPI Auflösung (z.B. 40/80 DPI) für die Paperbackausgabe und eine mittlere Auflösung für den Bilderzeugenden Drucktreiber (z.B. 200 DPI) verwendet wurde. Bei höherer Speicherdichte (> 100 DPI bei der Paperbackausgabe) wurden die erzeugten Images durch das verkleinernde Schreiben auf Mikrofilm leider im Rahmen der Rückdigitalisierung „unlesbar“. Dies ist neben der Vergrößerung auch durch den Qualitätsverlust bei der Rückdigitalisierung durch den Mikrofilmscanner bedingt. Auch bei hohen Auflösungen für die Digitalisierung tritt dieses Verhalten auf. Auch die rückdigitalisierten Optar Bilder lassen ein ähnliches Problem erwarten.

Es wurden die Probleme eines solchen Verfahrens bei der Nutzung der verwendeten Standardsoftware für das Mikrofilmschreiben und das Rückdigitalisieren des Mikrofilm eindeutig aufgezeigt (Skalierung beim Schreiben und suboptimales Rückdigitalisieren).

4 Ausblick

Zur Ablage in einem textbasierenden Format sollte ein eigenes Format entwickelt werden, welches auf die Problematiken bei der OCR Wandlung eingeht (d.h., Fehler bei der Erkennung von Zeichen wie z.B. 1 und I, tf und tt oder 0 und O) und zusätzlich eine Redundanz der abgelegten Daten und eine Fehlerkorrektur bietet. Auch der Einsatz einer entsprechenden geeigneten Schriftart (z.B. OCR A, OCR B) in einer entsprechend optimierten Größe können die Ergebnisse verbessern. Dabei ist zu berücksichtigen, dass die Wandlung in ein Textbasierendes Format die Anzahl der benötigten Seiten entsprechend erhöht und so auch einen entsprechenden Platzverbrauch auf dem Mikrofilm verursacht. Bei den barcode-artigen Formaten sind erste positive Ergebnisse mit Einschränkungen erzielt worden. Eine Optimierung bei der Erzeugung der Barcode-artigen Ablage, beim Schreiben der Daten auf Mikrofilm (skaliertes vs. Unskaliertes Schreiben), bei der Rückdigitalisierung (Vergrößerungsfaktor, Bildoptimierungen, …), bei der Verwendung von Redundanzen und der Fehlerkorrektur der abgelegten Daten in einem speziellen Barcode-artigem Format lassen Raum für weitere Untersuchungen, die zu einer interessanten und langzeitarchivierungstauglichen Lösung führen könnten.

Literaturverzeichnis

[BI06] „Bits on Film“, VDI VDE IT - InnoNet: Sichere Langzeitspeicherung "Bits on Film" (MILLENIUM), http://www.vdivde-it.de/innonet/projekte/k-o, PDF auf www.vdivde-it.de/innonet/projekte/in_pp146_millenium.pdf, 2006, Letzter Zugriff 13.03.2008

[BSIa] Bundesamt für Sicherheit in der Informationstechnik: IT- Grundschutz- Kataloge Punkt M 4.170 Auswahl geeigneter Datenformate für die Archivierung von Dokumenten; http://www.bsi.de/gshb/deutsch/m/m04170.htm; Stand 2005; Letzter Zugriff 23.3.2008

[BSIb] Bundesamt für Sicherheit in der Informationstechnik: IT- Grundschutz- Kataloge Punkt M 2.266 Regelmäßige Erneuerung technischer Archivsystem-Komponenten; http://www.bsi.de/gshb/deutsch/m/m02266.htm; Stand 2005; Letzter Zugriff 23.3.2008

[BSIc] Bundesamt für Sicherheit in der Informationstechnik: IT- Grundschutz- Kataloge Punkt G 2.72 Unzureichende Migration von Archivsystemen; http://www.bsi.de/gshb/deutsch/g/g02072.htm; Stand 2005; Letzter Zugriff 23.3.2008

[BSId] Bundesamt für Sicherheit in der Informationstechnik: IT- Grundschutz- Kataloge Punkt G 2.78 Unzulängliche Auffrischung von Datenbeständen bei der Archivierung; http://www.bsi.de/gshb/deutsch/g/g02078.htm; Stand 2005; Letzter Zugriff 23.3.2008

[RS60] Reed, I.S., Solomon, G.: "Polynomial codes over certain finite field," J. Soc. Ind. Applied Math., vol.8, pp.300–304, June 1960.

[WS07] Wendel, K; Schwitin, W: Schlußbericht zum Verbundprojekt ARCHE „Entwicklung eines Farbmikrofilm-Laserbelichters zur Langzeitarchivierung digitaler bzw. digitalisierter Dokumente“, Teilvorhaben „Entwicklung eines durchgehenden Workflow für die Erstellung von Farbmikrofilmen“, 13. März 2007 (ARCHE_v1.6_25.01.2007.pdf)

[Go49] Golay, M.J.E: Notes on digital coding. Proceedings of the IRE, 37:657, Juni 1949