H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 17.3

Bilddokumente

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

URL für Kapitel 17.3 „Bilddokumente“ (Version 2.3): urn:nbn:de:0008-20100305327
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100305327

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

17.3 Bilddokumente

Markus Enders

Digitale Bilddokumente (auch Images genannt) sind seid einigen Jahrzehnten in Gebrauch. Digitale Fotos, gescannte Dokumente oder anderweitig digital erzeugte Bilddokumente sind und werden millionenfach erstellt. Gedächtnisorganisationen müssen mit diesen Daten umgehen und sie langfristig archivieren können. Diese Aufgabe bietet verschiedene Herausforderungen. Unterschiedliche Datenformate, deren Benutzung und Unterstützung durch Softwareapplikationen bestimmten Moden unterliegen, sind nur ein Problem. Es stellt sich ferner die Frage, welche Metadata für Bilddokumente generiert werden können, wo diese gespeichert werden sollten und mit welchen Hilfsmitteln diese erzeugt bzw. extrahiert werden.

Seitdem Anfang der 1990er Jahre Flachbettscanner nach und nach in die Büros und seit Ende der 1990er Jahre auch zunehmend in die Privathaushalte einzogen, hat sich die Anzahl digitaler Bilder vervielfacht. Diese Entwicklung setzte sich mit dem Aufkommen digitaler Fotoapparate fort und führte spätestens seit der Integration kleiner Kameramodule in Mobiltelefone und Organizer sowie entsprechender Consumer-Digitalkameras zu einem Massenmarkt.
Heute ist es für Privatleute in fast allen Situationen möglich, digitale Images zu erzeugen und diese zu verschiedenen Zwecken weiterzubearbeiten. Der Markt bietet unterschiedliche Geräte an: von kleinen Kompaktkameras bis zu hochwertigen Scanbacks werden unterschiedliche Qualitätsbedürfnisse befriedigt.

Entsprechend haben sich auch Softwarehersteller auf diesen Markt eingestellt. Um Bilddokumente nicht im Dateisystem eines Rechners verwalten zu müssen, existieren heute unterschiedliche Bildverwaltungsprogramme für Einsteiger bis hin zum Profifotografen.

Diese Entwicklung kommt auch den Gedächtnisorganisationen zugute. Vergleichsweise günstige Preise ermöglichen es ihnen, ihre alten, analogen Materialien mittels spezieller Gerätschaften wie bspw. Scanbacks, Buch- oder Microfilmscannern zu digitalisieren und als digitales Image zu speichern. Auch wenn Texterfassungsverfahren über die Jahre besser geworden sind, so gilt die Authentizität eines Images immer noch als höher, da Erkennungs- und Erfassungsfehler weitestgehend ausgeschlossen werden können. Das Image gilt somit als „Digitales Master", von dem aus Derivate für Online-Präsentation oder Druck erstellt werden können oder deren Inhalt bspw. durch Texterkennung / Abschreiben für Suchmaschinen aufbereitet werden kann.

Datenformate

Digitale Daten müssen immer in einer für den Computer lesbaren und interpretierbaren Struktur abgelegt werden. Diese Struktur wird Datenformat genannt. Eine Struktur muss jedoch jeden Bildpunkt nicht einzeln speichern. Vielmehr können Bildpunkte so zusammengefasst werden, dass eine mehr oder weniger große Gruppe von Punkten als ein einzige Einheit gespeichert werden. Anstatt also jeden Bildpunkt einzeln zu speichern, belegen mehrere Bildpunkte denselben Speicherplatz. Die Art und Weise, wie diese Punkte zusammengefasst werden, wird als Komprimierungsalgorithmus bezeichnet. Dieser kann fest mit einem bestimmten Datenformat verbunden sein.

Sowohl das Datenformat als auch der Komprimierungsalgorithmus können bestimmte technische Beschränkungen haben. So kann durch das Format bspw. die Farbtiefe oder maximale Größe eines Bildes eingeschränkt sein. Der Komprimierungsalgorithmus kann bspw. nur auf reine schwarz-weiss Bilder angewendet werden.

In den letzten zwei Jahrzehnten wurde eine Vielzahl von Datenformaten für Bilddaten entwickelt. Zu Beginn der Entwicklung wirkten technische Faktoren stark limitierend. Formate wurden im Hinblick auf schnelle Implementierbarkeit, wenig Resourcenverbrauch und hohe Performanz während des Betriebs entwickelt. Dies führte zu vergleichsweise einfachen Lösungen, die auf einen bestimmten Anwendungszweck zugeschnitten waren. Teilweise wurden sie so proprietär, dass entsprechende Dateien nur von der Herstellersoftware, die zu einem Scanner mitgeliefert wurde, gelesen und geschrieben werden konnten. Der Austausch von Daten stand zu Beginn der Digitalisierung nicht im Vordergrund, so dass nur ein Teil der Daten zu Austauschzwecken in allgemein anerkannte und unterstützte Formate konvertiert wurden.

Heute ermöglicht das Internet einen Informationsaustausch, der ohne standardisierte Formate gar nicht denkbar wäre. Der Begriff „Standard“ ist aus Sicht der Gedächtnisorganisationen jedoch kritisch zu beurteilen, da „Standards“ häufig lediglich so genannte “De-facto“-Standards sind, die nicht von offiziellen Standartisierungsgremien erarbeitet und anerkannt wurden. Ferner können derartige Standards bzw. deren Unterstützung durch Hard- und Softwarehersteller lediglich eine kurze Lebenserwartung haben. Neue Forschungsergebnisse können schnell in neue Produkte und damit auch in neue Datenformate umgesetzt werden.

Für den Bereich der Bilddokumente sei hier die Ablösung des GIF-Formats durch PNG (Portable Network Graphics) beispielhaft genannt. Bis weit in die 1990er Jahre hinein war GIF der wesentliche Standard, um Grafiken im Inter-

net zu übertragen und auf Servern zu speichern. Dieses wurde aufgrund leistungsfähigerer Hardware, sowie rechtlicher Probleme durch das JPEG- und PNG-Format abgelöst. Heute wird das GIF-Format noch weitestgehend von jeder Software unterstützt, allerdings werden immer weniger Daten in diesem Format generiert. Eine Einstellung der GIF-Format-Unterstützung durch die Softwarehersteller scheint damit nur noch eine Frage der Zeit zu sein.

Ferner können neue Forschungsansätze und Algorithmen zu neuen Datenformaten führen. Forschungsergebnisse in dem Bereich der Wavelet-Komprimierung[7] sowie die Verfügbarkeit schnellerer Hardware führten bspw. zu der Erarbeitung und Implementierung des JPEG2000 Standards, der wesentlich bessere Komprimierungsraten bei besserer Qualität liefert als sein Vorgänger und zeigt, dass heute auch hohe Komprimierungsraten bei verlustfreier Komprimierung erreicht werden können.

Verlustfrei ist ein Komprimierungsverfahren immer dann, wenn sich aus dem komprimierten Datenstrom die Quelldatei bitgenau rekonstruieren lässt. Verlustbehaftete Komprimierungsverfahren dagegen können die Bildinformationen lediglich annäherungsweise identisch wiedergeben, wobei für das menschliche Auge Unterschiede kaum oder, je nach Anwendung, überhaupt nicht sichtbar sind.

Trotz eines starken Anstiegs der Übertragungsgeschwindigkeiten und Rechengeschwindigkeiten sind auch heute noch bestimmte Datenformate für spezifische Einsatzzwecke im Einsatz. Ein allgemeines Universalformat existiert nicht. Dies hat mitunter auch mit der Unterstützung dieser Formate durch gängige Internetprogramme wie Web-Browser, Email-Programme etc. zu tun. Nachfolgend sollen die gängigsten derzeit genutzten Datenformate kurz vorgestellt werden:

PNG (Portable Network Graphics): Dieses Datenformat wurde speziell für den Einsatz in Netzwerken entwickelt, um Bilder schnell zu übertragen und anzuzeigen. Entsprechend wurde ein Komprimierungsalgorithmus mit einem guten Kompromiss zwischen Dateigröße und Performanz gewählt. Dieser komprimiert das Bild verlustfrei. Überwiegend kommt dieses Format für die Anzeige von kleineren Images im Web-Browser zum Einsatz.

JPEG: Das JPEG Format komprimiert im Gegensatz zu PNG verlustbehaftet. D.h. das ursprüngliche Ergebnis-Bild lässt sich nach der Dekomprimierung

7 Weitere, einführende Informationen zu Wavelets finden sich unter: Graps, Amara (o.J.): An Introduction to Wavelets, http://www.amara.com/ftpstuff/IEEEwavelet.pdf

nicht mehr genau reproduzieren. Dadurch lässt sich ein wesentlich höherer Komprimierungsfaktor erreichen, der zu kleineren Dateien führt. Speziell für den Transfer von größeren Farbbildern in Netzwerken findet dieses Format Anwendung.

TIFF (Tagged Image File Format): TIFF wurde als universelles Austauschformat in 1980ern von Aldus (jetzt Adobe) entwickelt. Obwohl letzte Spezifikation zwar schon aus dem Jahr 1992 datiert,[8] ist es heute immer noch in Gebrauch. Dies liegt überwiegend an dem modularen Aufbau des Formats. Das Format definiert sogenannte Tags, die über eine Nummer typisiert sind. Entsprechend dieser Nummer enthalten diese Tags unterschiedliche Informationen. Somit ließen sich mit der Zeit neue Tags definieren, um neuartige Daten abzuspeichern. Auch die Art und Weise, wie die Bilddaten komprimiert werden ist nicht eindeutig definiert. Vielmehr definiert TIFF eine Liste unterschiedlicher Komprimierungsalgorithmen, die zum Einsatz kommen können. Darunter ist neben einigen verlustfreien Algorithmen auch dasselbe verlustbehaftete Verfahren zu finden, welches auch im JPEG Format angewandt wird. Als eines der wenigen Datenformate erlaubt TIFF auch die unkomprimierte Speicherung der Bilddaten. Aus diesem Grund wurde TIFF lange als einziges Format für die Speicherung der Archivversion eines digitalen Bildes (Master-Image) angesehen, auch wenn es nicht sehr effizient mit dem Speicherplatz umgeht. Dieser relativ große Speicherbedarf trug allerdings auch dazu bei, dass TIFF nicht als geeignetes Format für die Übertragung von Bilddaten im Internet angesehen wurde und mit der Entwicklung alternativer Formate wie GIF oder PNG begonnen wurde. Auch wenn bei heutigen Ressourcen und Bandbreiten dies nicht mehr ein so grosses Problem wäre, können TIFF-Dateien von keinem Web-Browser angezeigt werden.

JPEG2000: Ursprünglich wurde JPEG2000 als „Nachfolgeformat" für JPEG entwickelt. Hierbei wurde versucht Nachteile des JPEG Formats gegenüber TIFF unter Beibehaltung hoher Komprimierungsraten auszugleichen. Dies gelang durch die Anwendung neuartiger sogenannter Wavelet basierter Komprimierungsalgorithmen. Neben einer verlustbehafteten Komprimierung unterstützt JPEG2000 auch eine verlustfreie Komprimierung. Aufgrund des neuartigen Komprimierungsalgorithmus sind die erzeugten Dateien wesentlich kleiner als bei TIFF. Dies ist nicht zuletzt auch der Grund, warum JPEG2000

8 O.V.:TIFF 6.0 Specification. http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

neben TIFF als Datenformat für das „Digital Master“ eingesetzt wird, wenn es um das Speichern großer Farbbilder geht. Ähnlich des TIFF Formats können JPEG2000 Bilder derzeit nicht von einem Web-Browser angezeigt werden. Als Auslieferungsformat im Internet ist es daher derzeit nicht brauchbar.

Aus Perspektive der Langzeitarchivierung kommen also generell die Datenformate TIFF und JPEG2000 als Datenformat für das „Digital Master“ in Frage. Allerdings sind beide Formate so flexibel, dass diese Aussage spezifiziert werden muss.

Beide Formate können unterschiedliche Arten der Komprimierung nutzen. Diese ist entscheidend, um die Eignung als „Digital Master“-Format beurteilen zu können. So ist bspw. die LZW-Komprimierung für TIFF Images nach Bekanntwerden des entsprechenden Patents auf den Komprimierungsalgorithmus aus vielen Softwareprodukten verschwunden. Als Folge daraus lassen sich LZW-komprimierte TIFF Images nicht mit jeder Software einlesen, die TIFF unterstützt. Die Verlustbehaftete Komprimierung von JPEG2000 ist ebenfalls nicht als Format für das „Digital Master“ geeignet. Da hierbei Bytes nicht originalgetreu wieder hergestellt werden können, kommt für die Archivierung lediglich die verlustfreie Komprimierung des JPEG2000-Formats zum Einsatz.

Ferner spielt auch die Robustheit gegenüber Datenfehlern eine Rolle. So genannter „bitrot“ tritt mit der Zeit in fast allen Speichersystemen auf. Das bedeutet das einzelne Bits im Datenstrom kippen – aus der digitalen „1“ wird also eine „0“ oder umgekehrt. Solche Fehler führen dazu, dass Bilddateien gar nicht oder nur teilweise angezeigt werden können. Verschiedene Komprimierungsalgorithmen können entsprechend anfällig für einen solchen „bitrot“ sein. Datenformate können auch zusätzliche Informationen enthalten (sogenannte Checksums), um solche Fehler aufzuspüren oder gar zu korrigieren.

JPEG2000 bietet aufgrund seiner internen Struktur und des verwendeten Algorithmus einen weitreichenden Schutz gegen „bitrot“. Eine Fehlerrate von 0.01% im Bilddatenstrom (bezogen auf die Imagegesamtgröße) führt zu kaum sichtbaren Einschränkungen, wohingegen unkomprimierte TIFF-Dateien zu einzelnen fehlerhaften Zeilen führen können. Komprimierte TIFF-Dateien sind ohnehin wesentlich stärker von Bitfehlern beeinträchtigt, da der Bilddatenstrom nicht mehr vollständig dekomprimiert werden kann.[9]

---

9 Buonora, Paolo / Liberati, Franco: A Format for Digital Preservation – a study on JPEG 2000 File Robustness in: D-Lib Magazine, Volume 14, Number 7/8, http://www.dlib.org/dlib/july08/buonora/07buonora.html

Die Farbtiefe eines Bildes ist ebenfalls ein wichtiges Kriterium für die Auswahl des Datenformats für das „Digital Master". Rein bitonale Bilddaten (nur 1 bit pro Pixel, also reines Schwarz oder reines Weiß) können nicht im JPEG2000-Format gespeichert werden. Diese Bilddaten können jedoch im TIF-Format[10] durch die Verwendung des optionalen FaxG4-Komprimierungsalgorithmus sehr effizient gespeichert werden, welches verlustfrei komprimiert.

Den oben genannten Datenformaten ist gemein, dass sie von der Aufnahmequelle generiert werden müssen. Digitalkameras jedoch arbeiten intern mit einer eigenen an den CCD-Sensor angelehnten Datenstruktur. Dieser CCD-Sensor erkennt die einzelnen Farben in unterschiedlichen Sub-Pixeln, die nebeneinander liegen, wobei jedes dieser Sub-Pixel für eine andere Farbe zuständig ist. Um ein Image in ein gängiges Rasterimageformat generieren zu können, müssen diese Informationen aus den Sub-Pixeln zusammengeführt werden – d.h. entsprechende Farb-/Helligkeitswerte werden interpoliert. Je nach Aufbau und Form des CCD-Sensors finden unterschiedliche Algorithmen zur Berechnung des Rasterimages Anwendung. An dieser Stelle können aufgrund der unterschiedlichen Strukturen bereits bei einer Konvertierung in das Zielformat Qualitätsverluste entstehen. Daher geben hochwertige Digitalkameras in aller Regel das sogenannte „RAW-Format" aus, welches von vielen Fotografen als das Master-Imageformat betrachtet und somit archiviert wird. Dieses so genannte „Format" ist jedoch keinesfalls standardisiert.[11] Vielmehr hat jeder Kamerahersteller ein eigenes RAW-Format definiert. Für Gedächtnisinstitutionen ist diese Art der Imagedaten gerade über längere Zeiträume derzeit nur schwer zu archivieren. Daher wird zumeist auch immer eine TIFF- oder JPEG2000-Datei zusätzlich zu den RAW-Daten gespeichert.

Die Wahl eines passenden Dateiformats für die Images ist, gerade im Rahmen der Langzeitarchivierung, also relativ schwierig. Es muss damit gerechnet werden, dass Formate permanent auf ihre Aktualität, d.h. auf ihre Unterstützung durch Softwareprodukte, sowie auf ihre tatsächliche Nutzung hin überprüft werden müssen. Es kann davon ausgegangen werden, dass Imagedaten von Zeit zu Zeit in neue Formate überführt werden müssen, wobei unter Umständen auch ein Qualitätsverlust in Kauf genommen werden muss.

---

10 TIFF-Image oder TIFF-Datei aber TIF-Format, da in TIFF bereits „Format" enthalten ist (Tagged Image File Format).

11 Zu den Standarisierungsbestrebungen siehe http://www.openraw.org/info sowie http://www.adobe.com/products/dng/

Metadaten für die Archivierung

Ziel der Langzeitarchivierung ist das dauerhafte Speichern der Informationen, die in den Bilddokumenten abgelegt sind. Das bedeutet nicht zwangsläufig, dass die Datei als solche über einen langen Zeitraum aufbewahrt werden muss. Es kann bspw. erforderlich werden Inhalte in neue Formate zu überführen. Eine sogenannte Migration ist immer dann erforderlich, wenn das Risiko zu hoch wird ein bestimmtes Datenformat nicht mehr interpretieren zu können, weil kaum geeignete Soft- oder Hardware zur Verfügung steht.

Neben dem dauerhaften Speichern der Bilddaten ist es ebenfalls wichtig den Kontext der Bilddaten zu sichern. Unter Kontext sind in diesem Fall alle Informationen zu verstehen, die den Inhalt des Bilddokuments erst zu- und einordnen lassen. Dies ist in aller Regel der Archivierungsgegenstand. So ist bspw. eine einzelne als Bild digitalisierte Buchseite ohne den Kontext des Buches (= Archivierungsgegenstand) nicht einzuordnen. Im dem Fall eines Katastrophenszenarios, in dem auf zusätzliche Informationen, wie sie in etwa ein Repository oder ein Katalog enthält, nicht zugegriffen werden kann, weil entweder das System nicht mehr existiert oder aber die Verknüpfung zwischen System und Bilddokument verloren gegangen ist, können zusätzliche Metadaten, die in dem Bilddokument direkt gespeichert werden, den Kontext grob wieder herstellen.

Deskriptive Metadaten in Bilddokumenten

Diese sogenannten deskriptiven Metadaten, die den Archivierungsgegenstand und nicht das einzelne Bilddokument beschreiben, können direkt in jedem Bilddokument gespeichert werden. Jedes Datenformat bietet dazu eigene proprietäre Möglichkeiten.

Frühe Digitalisierungsaktivitäten haben dazu bspw. die TIFF-Tags PAGENAME, DOCUMENTNAME und IMAGEDESCRIPTION genutzt, um entsprechende deskriptive Metadaten wie Titelinformation und Seitenzahl abzubilden.[12] Diese sind mitunter auch heute noch in Digitalisierungsprojekten gebräuchlich. Eine weniger proprietäre Lösung ist die von Adobe entwickelte Extensible Metadata Plattform (XMP).[13] Zum Speichern von deskriptiven

12 O.V.: Bericht der Arbeitsgruppe Technik zur Vorbereitung des Programms „Retrospektive Digitalisierung von Bibliotheksbeständen“ im Förderbereich „Verteilte Digitale Forschungsbibliothek“, Anlage 1, http://www.sub.uni-goettingen.de/ebene_2/vdf/anlage1.htm

13 O.V.: XMP Specification, September 2005, http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf

Metadaten verwendet XMP das Dublin Core Schema. XMP-Daten können sowohl zu TIFF und JPEG2000 hinzugefügt werden als auch zu PDF und dem von Adobe entwickeltem Bilddatenformat für RAW-Bilddaten DNG.

Im Falle eines Katastrophenszenarios im Rahmen der Langzeitarchivierung lässt sich mittels dieser XMP-Daten ein entsprechender Kontext zu jedem Bilddokument wieder aufbauen.

Technische Metadaten für Bilddokumente

Jede Datei hat aufgrund ihrer Existenz inhärente technische Metadaten. Diese sind unabhängig vom verwendeten Datenformat und dienen bspw. dazu die Authentizität eines Images zu beurteilen. Checksummen sowie Größeninformationen können Hinweise darauf geben, ob ein Image im Langzeitarchiv modifiziert wurde.

Darüber hinaus gibt es formatspezifische Metadaten. Diese hängen direkt vom eingesetzten Datenformat ab und enthalten bspw. allgemeine Informationen über ein Bilddokument:

- Bildgröße in Pixel sowie Farbtiefe und Farbmodell
- Information über das Subformat – also bspw. Informationen zum angewandten Komprimierungsalgorithmus, damit der Datenstrom auch wieder entpackt und angezeigt werden kann.

Mittels Programmen wie bspw. JHOVE[14] lassen sich eine Vielzahl von technischen Daten aus einer Datei gewinnen. Gespeichert wird das Ergebnis als XML-Datei. Als solche können die Daten in Containerformate wie bspw. METS eingefügt und im Repository gespeichert werden. Aufgrund der Menge der auszugebenden Informationen sind diese allerdings kritisch zu bewerten. Entsprechende Datensätze bspw. für ein digitalisiertes Buch sind entsprechend groß. Daher wird nur in seltenen Fällen der komplette Datensatz gespeichert, sondern bestimmte technische Metadaten ausgewählt. Für Bilddokumente beschreibt NISO Z39.87 ein Metadatenschema für das Speichern von technischen Metadaten.[15] Eine entsprechende Implementierung in XML steht mit MIX ebenfalls bereit.[16]

---

14 JHOVE – JSTOR/Harvard Object Validation Enviroment, http://hul.harvard.edu/jhove/

15 O.V.: Data Dictionary – Technical Metadata for Digital Still Images, http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c

16 http://www.loc.gov/standards/mix/

Es ist anzunehmen, dass zukünftig Migrationsprozesse vor allem bestimmte Sub-Formate betreffen werden, also bspw. nur TIFF-Dateien mit LZW-Komprimierung anstatt alle TIFF-Dateien. Für die Selektion von entsprechenden Daten kommt dem Format also eine große Bedeutung zu. Mit PRONOM steht eine Datenbank bereit, die Dateiformate definiert und beschreibt. Dabei geht die Granularität der Datenbank weit über gängige Formatdefinitionen, wie sie bspw. durch den MIME[17]-Type definiert werden, hinaus. TIFF-Dateien mit unterschiedlicher Komprimierung werden von PRONOM[18] als unterschiedliche Formate verstanden. Um diese Formatinformationen aus den Bilddokumenten zu extrahieren steht mit DROID[19] ein entsprechendes Tool zur Verfügung.

Herkunftsmetadaten für Bilddokumente

Für die Langzeitarchivierung sind neben technischen Metadaten auch Informationen über die Herkunft der Bilddateien wichtig. Informationen zur eingesetzten Hard- und Softwareumgebung können hilfreich sein, um später bestimmte Gruppen zur Bearbeitung bzw. Migration (Formatkonvertierungen) auswählen oder aber um Bilddokumente überhaupt darstellen zu können.

Im klassischen Sinn werden Formatmigrationen zwar anhand des Dateiformats ausgewählt. Da jedoch Software selten fehlerfrei arbeitet, muss bereits bei der Vorbereitung der Imagedaten Vorsorge getroffen werden, entsprechende Dateigruppen einfach selektieren zu können, um später bspw. automatische Korrekturalgorithmen oder spezielle Konvertierungen durchführen zu können.

Ein nachvollziehbares und in der Vergangenheit real aufgetretenes Szenario ist bspw. die Produktion fehlerhafter PDF-Dateien auf Basis von Images durch den Einsatz einer Programmbibliothek, die sich im nachhinein als defekt erwies. In der Praxis werden diese nur zugekauft, sodass deren Interna dem Softwareanbieter des Endproduktes unbekannt sind. Tritt in einer solchen Programmbibliothek ein Fehler auf, so ist dieser eventuell für den Programmierer nicht auffindbar, wenn er seine selbst erzeugten Dateien nicht wieder einliest (bspw. weil Daten nur exportiert werden). Ein solcher Fehler kann auch nur in einer bestimmten Softwareumgebung (bspw. abhängig vom Betriebssystem) auftreten. Kritisch für die Langzeitarchivierung wird der Fall dann, wenn einige Softwareprodukte solche Daten unbeanstan-

17 Freed, N; Borenstein, N (1996): Multipurpose Internet Mail Extensions (MIME) part one, RFC2045, http://tools.ietf.org/html/rfc2045

18 http://www.nationalarchives.gov.uk/pronom/

19 Digital Record Object Identification (DROID): http://droid.sourceforge.net

det laden und anzeigen, wie in diesem Fall der Adobe PDF-Reader. „Schwierigkeiten“ hatten dagegen OpenSource Programme wie Ghostscript sowie die eingebauten Postscript-Interpreter einiger getesteter Laserdrucker.

Trotz gewissenhafter Datengenerierung und Überprüfung der Ergebnisse kann es also dazu kommen, dass nicht konforme Bilddokumente über Monate oder Jahre hinweg produziert werden. Entsprechende Informationen zur technischen Laufzeitumgebung erleichtern jedoch die spätere Identifikation dieser „defekten“ Daten im Langzeitarchivierungssystem.

Eine weitere Aufgabe der Herkunftsmetadaten ist es den Lebenszyklus eines Dokuments aufzuzeichnen. Durch Verweise auf Vorgängerdateien können Migrationsprozesse zurückverfolgt werden. Dies gewährleistet, dass auch auf frühere Generationen als Basis für eine Migration zurückgegriffen werden kann. Im Fall von „defekten“ Daten ist das eine wesentliche Voraussetzung, um überhaupt wieder valide Inhalte generieren zu können.

Sowohl technische als auch Herkunftsmetadaten werden als eigenständige Metadatenrecords unter Verwendung spezifischer Metadatenschemata gespeichert. Für Bilddokumente bietet sich MIX für die technischen Metadaten an. Da Herkunftsmetadaten nicht spezifisch auf Bilddokumente zugeschnitten sind, stellen allgemeine Langzeitarchivierungsmetadatenschemata wie bspw. PREMIS[20] entsprechende Felder bereit.

Um die unterschiedlichen Metadaten zusammen zu halten, kommt darüber hinaus ein Containerformat wie METS[21] oder MPEG-21 DIDL[22] zum Einsatz.

Ausblick

Sollen Bilddokumente entsprechend der oben skizzierten Anforderungen für die Langzeitarchivierung vorbereitet werden, ist es aus praktischer Sicht unerlässlich aktuelle Werkzeuge und Geschäftsprozesse zu evaluieren. Viele Werkzeuge sind bspw. nicht in der Lage entsprechende Metadaten wie bspw. XMP in einem Bilddokument zu belassen. Ein Speichern des Bilddokuments sichert zwar den entsprechenden Bilddatenstrom, lässt die deskriptiven Metadaten außen vor.

---

20 http://www.loc.gov/premis

21 Siehe Kap. 6.2 Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten

22 Bekart, Jeroen; Hochstenbach, Patrick; Van de Sompel Herbert (2003): Using MPEG-21 DIDL to represent complex objects in the Los Alamos National Laboratory Digital Library In: D-Lib Magazine, Band 9, November 2003, http://igitur-archive.library.uu.nl/DARLIN/2005-0526-201749/VandeSompelDLib2003UsingMPEG.htm

Das Vorbereiten der Bilddokumente für die Langzeitarchivierung ist in aller Regel ein mehrstufiger Prozess. Dieser Prozess muss wohl dokumentiert und gesteuert werden, um eine gleichbleibende Qualität sicherzustellen. Ein „spontan" durchgeführtes Laden und Abspeichern eines Images könnte dazu führen, dass sich bspw. technische Metadaten wie die Checksumme ändern, da eigene, zusätzliche Metadaten durch die Software eingefügt wurden. In der Praxis hat sich für die Aufbereitung von Bilddokumente folgender, hier stark vereinfachter Workflow als sinnvoll erwiesen:

- Einfügen der deskriptiven Metadaten in das Bilddokument
- Validieren des Datenformates des Bilddokuments
- Extrahieren der Formatinformation (JHOVE) inkl. der Formatbestimmung (DROID)
- Extrahieren der allgemeinen technischen Metadaten (Checksummen)
- Generierung der technischen und Herkunftsmetadaten (MIX und PREMIS) aus den Formatinformationen
- Einfügen der technischen und Herkunftsmetadaten in ein Containerformat des Repositories.

Aufgrund der Menge an Bilddokumenten ist dieser Prozeß nur automatisiert durchführbar. Um Fehler zu vermeiden und auch auf nachträglich notwendige Korrekturen reagieren zu können, ist der Einsatz spezieller Software zur Steuerung von Geschäftsprozessen sinnvoll. Dadurch wird eine gleichbleibende Qualität gewährleistet. Ferner ist zu hoffen, dass damit Zeitaufwand und Kosten für die Langzeitarchivierung von Bilddokumenten sinken.