nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

15.2 Bilddokumente

Herausgeber:

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Im Auftrag von:

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:

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de
Tel. +49 (0) 55 1 39 38 66

15.2 Bilddokumente

Markus Enders

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 Zwecke weiterzubearbeiten. Der Markt bietet unterschiedliche Geräte zu unterschiedlichen Zwecken 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 Gedächnisorganisationen zu gute. 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

Die seit über zwei Jahrzehnten statt findende Digitalisierung von Bildmaterialien hat zu einer Vielzahl unterschiedlicher Datenformate geführt. Gerade zu Beginn waren die technischen Faktoren limitierend, was aus Gründen schneller Implementierbarkeit und einfachen Handlings während des Betriebs zu „einfachen" technischen Lösungen führte. Diese waren teilweise so proprietär, dass sie nur von der Herstellersoftware gelesen und geschrieben werden konnten. Datenaustausch stand zu Beginn der Digitalisierung nicht im Vordergrund, sodass 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 Gedächnisorganisationen jedoch kritisch zu beurteilen, da „Standards“ häufig lediglich so genannte “De-facto“-Standards sind, die nicht von offiziellen Standarisierungsgremien 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 Internet zu übertragen und auf Servern zu speichern. Dieser wurde aufgrund leistungsfähigerer Hardware, sowie rechtlicher Probleme durch das JPEG- und PNG-Format abgelöst. Heute wird das GIF-Format noch weitestgehend unterstützt, allerdings werden immer weniger Daten in diesem Format generiert. Eine Einstellung der GIF-Format-Unterstützung durch die Softwarehersteller ist damit nur noch eine Frage der Zeit.
Ferner können neue Forschungsansätze und Algorithmen zu neuen Datenformaten führen. Forschungsergebnisse in dem Bereich der Wavelet-Komprimierung 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.
Neben dem oben erwähnten JPEG- oder PNG-Format, findet heute vor allem das TIFF-Format für die Master-Images Einsatz. Dessen Spezifikation beschreibt allerdings lediglich den prinzipiellen Aufbau dieses Formats und lässt viel Raum für eigene Erweiterungen und Komprimierungsmethoden. Daher lohnt sich ein genauer Blick darauf, welche Komprimierungsmethoden von einer Software unterstützt werden und welche Risiken mit deren Nutzung verbunden sind. So ist bspw. die LZW-Komprimierung für TIFF Images nach

Bekannt werden 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.
Aus Sicht der Langzeitarchivierung ist daher heute Stand der Technik die Nutzung des unkomprimierten TIFF-Formats für Graustufen- und Farbimages. Dies ist jedoch aufgrund des Platzbedarfs gerade für hochaufgelöste Images recht umstritten. Als Nachfolger wird derzeit der JPEG2000-Standard gehandelt, der vor allem in seiner verlustfreien Variante, dieselbe Qualität erreicht, jedoch wesentlich weniger Speicherplatz einnimmt. Derzeit behindert die mangelnde Unterstützung seitens der Softwarehersteller die Einsatzfähigkeit des neuen Formates: viele Programme können JPEG2000 nicht lesen oder schreiben, obwohl mittlerweile kostengünstige Programmierlibraries sowie kleine Konvertierungstools auf dem Markt sind.
Für reine schwarz-weiß (bitonale) Images hat sich die FaxG4-Komprimierung bewährt, da sie sehr gute Komprimierungsraten erlaubt und verlustfrei arbeitet.
Den oben genannten Dateiformaten 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 einem gängigen Rasterimageformat generieren zu können, müssen diese Information 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[7]. Vielmehr hat jeder Kamerahersteller ein eigenes RAW-Format definiert. Für Gedächnisinstitutionen 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

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

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.

Bedeutung der Metadaten für die Archivierung

Die Speicherung und Lagerung der Imagedaten über längere Zeiträume kann dazu führen, dass Daten nur noch teilweise lesbar sind. Ebenfalls können Daten durch fehlerhafte Konvertierungsprozesse zerstört werden. Die Probleme, die zu bewältigen sind, sind vielfältig:
Während der Lagerung und Konvertierung von Daten ist der Kontext einzelner Images beizubehalten. D.h. die Zugehörigkeit einzelner Seiten oder anderer digitalisierter Objekte zu einem größeren Kontext (bspw. eines Buches) muss sichergestellt werden. Dazu ist es gerade für die Langzeitarchivierung ratsam, entsprechende Daten zusätzlich zu externen Metadatensätzen auch direkt im jeweiligen Image unterzubringen. Das TIFF-Dateiformat kennt dazu bzw. die TIFF-Tags PAGENAME, DOCUMENTNAME und IMAGEDESCRIPTION, um Informationen zu dem jeweiligen Image zu speichern. Da es sich um freie Textfelder handelt, ist prinzipiell das Abspeichern von XML-Strukturen innerhalb des Feldes möglich.

- PAGENAME kann bspw. die jeweilige Seitennummer innerhalb des Buches enthalten. Auch wenn bspw. der Dateiname eines Images verloren geht, kann immer die Reihenfolge der verschiedenen Imagedateien innerhalb des übergeordneten Kontexts bestimmt werden.
- DOCUMENTNAME sollte Informationen zum übergeordneten Kontext enthalten, die diesen eindeutig identifizieren. Dies kann der Titel, der Autor oder aber auch der Identifier (bspw. die ISBN oder eine Katalognummer) sein.
- IMAGEDESCRIPTION kann weiterführende Informationen zum Kontext des Images enthalten, bspw. die komplette bibliographische Information.

Für die Langzeitarchivierung sind auch Metadaten hinsichtlich der Generierung sowie des Generierungsprozesses wichtig. Informationen zur eingesetzten Hard- und Softwareumgebung hilfreich sein, um später bestimmte Gruppen zur Bearbeitung bzw. Migration (Formatkonvertierungen) auswählen 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 mittels einer defekten Programmbibliothek. Diese so genannten „Libraries" werden von verschiedenen Softwareherstellern häufig nur zugekauft, sodass deren Interna ihnen unbekannt sind. Tritt in dieser Programmbibliothek ein Fehler auf, so ist dieser eventuell für den Programmierer nicht auffindbar, da er seine selbst erzeugten Dateien nicht wieder einliest. Dies gilt vor allem für klassische Exportfunktionen.
In dem oben erwähnten Szenario erzeugte die entsprechende Programmbibliothek nur unter dem Solaris Betriebssystem fehlerhafte PDF-Dateien, bei denen ein „." (Punkt) durch ein „," ersetzt wurde. Kritisch für die Langzeitarchivierung wird der Fall dann, wenn einige Softwareprodukte solche Daten unbeanstandet laden und anzeigen, wie in diesem Fall der Adobe PDF-Reader. „Schwierigkeiten" machten dagegen OpenSource Programme wie Ghostscript sowie die eingebauten Postscript-Interpreter einiger getesteter Laserdrucker.
Letztlich kann dies dazu führen, dass solche Daten über Monate oder Jahre hinweg produziert werden. Werden entsprechende Informationen zur technischen Laufzeitumgebung zu jedem einzelnen Image gespeichert, kann das Data-Management eines Langzeitarchivierungssystem entsprechende Dateien identifizieren und für eine Fehlerbehebung selektieren.
Die besondere Gefahr bei der Be- und Verarbeitung dieser so genannten „Embedded"-Metadaten besteht darin, dass sie, obwohl von Standards vorgesehen häufig nicht durch entsprechende Implementierungen berücksichtigt werden. D.h. diese Metadaten gehen häufig beim Speichern nach einem Bearbeitungsschritt verloren. Für die Langzeitarchivierung bedeutet dies, dass diese Metadaten direkt vor dem Einspielen in das Langzeitarchivierungssystem überprüft und ggf. erzeugt werden müssen.

Technische Metadaten für Imagedateien

Jede Datei hat aufgrund ihrer Existenz technische Metadaten. Dies sind so genannte formatunabhängige Metadaten, die u.a. auch dazu dienen können 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 Dateiformat ab und enthalten bspw. allgemeine Informationen über ein Image:

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

Diese Daten lassen sich direkt aus einer Imagedatei gewinnen. Das Tool JHOVE ist bspw. in der Lage diese Daten zu erzeugen und als XML-Datei auszugeben. Im Rahmen der Langzeitarchivierung können diese Informationen sinnvoll bspw. zur Selektion von Daten verwendet werden, indem Migrationsprozesse abhängig von der jeweiligen Farbtiefe andere Zielformate definieren.

Generierungsprozess von Images

Um Images zu Einheiten zu gruppieren und diese entsprechend mit Metadaten zu beschreiben, ist der Einsatz von so genannten Containerformaten sinnvoll. Diese beschreiben ein komplexes Objekt, welches durch ein oder mehrere Images wiedergegeben wird. Diese Daten sind nicht innerhalb der Images gespeichert, sondern liegen teilweise redundant außerhalb des digitalen Objekts vor.

Ein entsprechendes Containerformat, das ein Archivsystem zum Einspielen der Images benötigt, könnte bspw. METS oder MPEG-21 DIDL sein.

Die Information zum Kontext sowie die entsprechenden Metadaten können selten sinnvoll in einem Arbeitsschritt erfasst werden. Vielmehr ist der Einsatz spezieller Software zur Steuerung von Geschäftsprozessen sinnvoll, die diese Daten erfasst, den einzelnen Images zuordnet und anschließend zusammen als ein Paket mit den Images an das Langzeitarchivierungssystem überführt. Werden Informationen zu einzelnen Arbeitsschritten erfasst, ist nachträglich auch die Beurteilung der Imagequalität im Langzeitarchiv möglich, da bspw. entsprechende Be- und Verarbeitungsmaßnahmen Rückschlüsse auf die ursprünglich erzeugte Imagedatei zulassen.

Ausblick auf die Aufbereitung von Imagedaten zur Langzeitarchivierung

Da es heute für die Generierung und Speicherung von Imagedaten bewährte Technologien gibt, hängt die Möglichkeit Bilddokumente langfristig zu archivieren und in einer ihrem Originalzustand entsprechenden oder weitgehend angenäherten Qualität verfügbar zu machen, von der Berücksichtigung der o. g. Faktoren ab.

Generell lässt sich sagen, dass eine genaue Planung und Dokumentation hinsichtlich eingesetzter Software, benutzter Formate und erfasster Metadaten diese Aufgabe vereinfachen wird. Ferner wird zukünftig Software zur Verwaltung und Steuerung von Geschäftsprozessen, gerade bei der Generierung von Bilddokumenten, die Kosten zur Erfassung dieser zusätzlichen Informationen senken. Nicht zuletzt deswegen ist zu hoffen, dass die Kosten für die Langzeitarchivierung von Bilddokumente sinken, auch wenn deren Produktionskosten zunächst leicht ansteigen werden.