Chancen und Risiken verlustbehafteter Bildkompression in der digitalen Archivierung

Kai Naumann, Christoph Schmidt

Einleitung

Zu den Kernanliegen des archivischen Berufs gehört es, das dem Archiv anvertraute Archivgut vor Beschädigungen und Verlust zu bewahren. Es ist daher nicht verwunderlich, dass die willentliche Herstellung oder Duldung von Verlust in der Archivwelt nur selten ergebnisoffen diskutiert wird. Dies gilt auch und insbesondere für die Nutzung so genannter «verlustbehafteter» Bilddatenformate bei der digitalen Archivierung. Zumindest in der deutschen Archivcommunity besteht heute ein stillschweigendes Übereinkommen, den Einsatz entsprechender Techniken mehr oder minder explizit abzulehnen.[1] Gleichwohl prägen verlustbehaftete Bilddatenformate die außerarchivische digitale Landschaft in hohem Maße. So hat sich etwa JPEG in den vergangenen 20 Jahren zu einem Bilddatenformat entwickelt, das aus vielen Anwendungsgebieten gar nicht mehr fortzudenken scheint. Insbesondere im Kontext des Internet, aber auch in der Digitalfotografie ist JPEG omnipräsent, so dass die Archive in wachsendem Maße gezwungen sind, sich mit Anbietungen auseinanderzusetzen, die verlustbehaftete Bilddatenformate enthalten. Es ist daher nicht verwunderlich, dass das archivische Ablehnungsdogma ebenso langsam wie stillschweigend erodiert. So speichern viele Archive inzwischen Bilder im JPEG-Format, sofern dieses gleichzeitig auch das Produktions- bzw. Anbietungsformat ist. Eine mögliche Neubewertung verlustbehafteter Bilddatenformate ergibt sich jedoch auch in der Digitalisierung aus den wachsenden Bilddatenmengen, die in einigen Bereichen dazu führen, das Kriterium der ökonomischen Effizienz bei der Auswahl geeigneter Speicherformate vorsichtig neu zu gewichten.[2]

1 Diese Ablehnung spiegelt sich auch in verschiedenen Empfehlungen und Vorschriften wider, wie etwa den «Praxisregeln Digitalisierung» der Deutschen Forschungsgemeinschaft (Deutsche Forschungsgemeinschaft: DFG-Praxisregeln «Digitalisierung». DFG-Vordruck 12.151, o.O., [2016], S. 20-21) oder dem «Katalog archivischer Datenformate» der KOST (http://kost-ceco.ch/wiki/whelp/KaD/index.php). (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Als Beispiele wären hier etwa die besonders in den größeren Flächenländern sehr umfangreichen Luftbilddatenbestände der staatlichen Vermessungsverwaltungen zu nennen. In den von der Vermessungsverwaltung und den Archiven gemeinsam erstellten «Leitlinien zur bundesweit einheitlichen Archivierung von Geobasisdaten» wird der Einsatz verlustbehafteter Bildkompressionsverfahren zumindest nicht strikt abgelehnt (Leitlinien zur bundesweit einheitlichen Archivierung von Geo-

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.7

Vor diesem Hintergrund ist es das Ziel des vorliegenden Textes, einen ergebnisoffenen Dialog über verlustbehaftete Speicherformate anzustoßen und zu fördern. Da das gesamte Thema viele unterschiedliche Teilaspekte hat, die jeweils für sich genommen bereits eine intensivere Behandlung verdienten, wäre es im Rahmen des gewählten Publikationsformats unmöglich, diese erschöpfend zu diskutieren. Der Text beschränkt sich daher darauf, Anregungen und Denkanstöße zu bieten, Fragen aufzuwerfen und Vorschläge zu machen.

Am Anfang unserer Überlegungen stehen einige allgemeine Gedanken zum Verlustbegriff, der, wie zu zeigen sein wird, eng mit dem weniger negativ konnotierten Begriff der «Veränderung» verbunden ist. Danach soll anhand einiger Beispiele beleuchtet werden, welche Verluste beim Einsatz verlustbehafteter Kompressionsverfahren unter Laborbedingungen auftreten und wie diese zu bewerten sind. Unsere Darstellungen basieren dabei auf einigen praktischen Versuchen, die wir im Vorfeld der AUdS-Tagung 2017 unternommen haben.[3] Diese haben, darauf sei ausdrücklich hingewiesen, nicht den Anspruch wissenschaftlicher Beweisführung; sie sollen vielmehr zu eigenen Gedanken, Experimenten, Widersprüchen anregen. Nach der Vorstellung einiger praktischer Facetten im Umgang mit verlustbehafteten Kompressionsverfahren sollen dann einige Argumente diskutiert werden, die für den Einsatz verlustbehafteter Formate sprechen. Der Fokus der Überlegungen liegt dabei auf dem am weitesten verbreiteten Format JPEG. Ein knappes Fazit fasst die wesentlichen Ergebnisse des Beitrags zusammen.

Der Verlustbegriff in der digitalen Bestandserhaltung

Die Reprographie gehört zu den Disziplinen, die unter bestimmten Bedingungen Verluste in Kauf nehmen können, denn reprographische Verluste sind zunächst nur Qualitätsveränderungen, ohne dass damit eine Verschlechterung des angestrebten operativen Ergebnisses verbunden sein muss. Veränderungen können nämlich durchaus im Sinne des Erstellers sein, wenn etwa eine Wandlung in JPEG zwar im Detail zu Veränderungen führt, aber das Verschicken eines großen Digitalfotos per E-Mail erlaubt. Auch ein Archiv, das ja gleichsam eine Datei viele 100 Jahre in die Zukunft verschicken soll, kann sich überlegen, welche Qualitätsveränderungen erlaubt und welche unerwünscht sind.[4]

basisdaten. Abschlussbericht der gemeinsamen AdV-KLA-Arbeitsgruppe «Archivierung von Geobasisdaten» 2014-2015, o.O., [2015], S. 15-16).

3 Diese Versuche wurden von einigen Kollegen in den beteiligten Archiven tatkräftig unterstützt. Der Dank der Autoren hierfür gilt vor allem Corinna Knobloch (Landesarchiv Baden-Württemberg) und Martin Hoppenheit und Marcel Werner (Landesarchiv Nordrhein-Westfalen).

4 So auch der Fototechnische Ausschuss der Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA) in seinem Papier von Dezember 2016, allerdings ohne

Wer die Grenzen des Erlaubten definieren will, kann bei der Umwandlung von Rastergrafiken (Pixelgrafiken) in Folgeformate absolute Maßstäbe setzen. Jedes Pixel, das kann die Anforderung sein, soll seine Position im Bild und seine Farbe behalten. Dazu müssen stets auch das Farbmodell (z.B. RGB) und das Farbprofil (z.B. sRGB) erhalten bleiben. Das Ergebnis nennt man eine verlustfreie Formatmigration. Doch es können auch relative Maßstäbe gesetzt werden, die Veränderungen wie die berühmten JPEG-Kompressionsartefakte in gewissen Grenzen erlauben. Gibt es unter den relativen Maßstäben auch für das Archiv hinnehmbare Varianten?

In jedem Fall ist festzustellen, dass in den letzten Jahren für die Archive die Möglichkeiten zugenommen haben, Verluste sichtbar zu machen und sie zu bewerten. Das kostenlose Open-Source-Programmpaket ImageMagick steht bereit, um selbst weniger gebräuchliche Formate wie JPEG2000 zu erstellen und zu analysieren.[5] Die KOST hat für ImageMagick außerdem eine grafische Oberfläche namens KOST-Simy bereitgestellt, die ein Rasterbild Pixel für Pixel mit einem anderen Rasterbild vergleicht, das die gleiche Anzahl Pixel in Höhe und Breite besitzt. Die Differenz zwischen beiden Bildern wird in einer absoluten Zahl an Pixeln, einer prozentualen Quote an Pixeln (in Relation zur Gesamtzahl) und einer Rastergrafik mit Falschfarben festgehalten. Ein Toleranzwert für den Farbunterschied je Pixel und die Differenzquote können eingestellt werden (vgl. unten Versuchsaufbau 1, 3. Absatz).

Diese Differenz als Quote veränderter Pixel ist für Formatmigrationen von Rastergrafiken eine messbare Eigenschaft, die grundsätzlich ins Konzept der digitalen Bestandserhaltung gemäß Nestor-Leitfaden «Digitale Bestandserhaltung»[6] passt. Die Differenzzahl ermöglicht Aussagen zum Erfüllungsgrad der dort genannten Kriterien E1, E4, E5, E6 und E7, aber nur in einem übergreifenden Sinne. Die Kriterien einzeln abzuprüfen, ist mit dieser Zahl nicht möglich. Im Leitfaden wurde für Bilder insgesamt, das heißt sowohl Vektorgrafik als auch Rastergrafik (oder Pixelgrafik), ein Anforderungskatalog formuliert. In den hier nachfolgend dargestellten Experimenten war festzustellen, in welchem Ausmaß die Differenz auftritt, wie sie beeinflussbar ist und ob die im Nestor-Leitfaden genannten Erfüllungsgrade ausreichen, um die Differenz zu beschreiben. Nicht möglich war es, die im Leitfaden geforderten Prüfungen zu vollziehen. Entsprechend würde es sich empfehlen, in der Weiterentwicklung des Leitfadens Rastergrafik und Vektorgrafik zu unterscheiden und auf durch Software prüfbare Eigenschaften noch mehr Wert zu legen.

Überlegungen zur Kompression, http://www.bundesarchiv.de/DE/Content/Downloads/KLA/wirtschaftliche-digitalisierung.pdf?__blob=publicationFile.

5 http://www.imagemagick.org/.

6 Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung, Version 2.0. Verfasst und herausgegeben von der nestor-Arbeitsgruppe Digitale Bestandserhaltung, Frankfurt am Main 2012. http://nbn-resolving.de/urn:nbn:de:0008-2012092400.

Beispiele für den Einsatz verlustbehafteter Bildkompression unter Laborbedingungen

Versuchsaufbau 1

Das erste Beispiel, das experimentell im Staatsarchiv Ludwigsburg bearbeitet wurde, ist ein JPEG-Bilddatenstrom in einer PDF/A-Datei. Der PDF/A-Standard erlaubt seit seiner ersten Ausprägung PDF/A-1 das Einbetten von JPEG-Objekten. Die meisten Kopierer erstellen, wenn man eine farbige Vorlage in PDF/A scannt, eine solche Datei. Es ist also sehr wahrscheinlich, dass sich solche JPEG-Bilddatenströme in Archivbeständen finden und eines Tages (zum Beispiel im Jahr 2056) in ein Folgeformat migriert werden müssen. Da JPEG eine sehr ökonomische Relation zwischen Datenmenge und Pixelanzahl besitzt, würde eine Formatmigration in ein verlustfreies Format (wie heute z.B. PNG oder TIFF) eine erhebliche Vermehrung der Datenmenge um den Faktor 2 bis 3 mit sich bringen. Sollten JPEG-Datenströme in der Größenordnung von 3 TB vorliegen, müssten für die neue Repräsentation nicht weitere 3 TB, sondern weitere 9 TB bereitgestellt werden. Möglich ist, dass Speicherkosten dann keine Rolle mehr spielen werden. Es erscheint aber auch möglich, dass eine finanzielle Erwägung zu der Vorgabe führt, dass eine neue Repräsentation ebenfalls nur 3 TB oder besser weniger verbrauchen darf. Und ebenso könnte es bei einer zweiten Migration im Jahr 2098 sein.

Um dieses Szenario zu simulieren, wurde ein Testbild in Repräsentation R 1 zunächst vom JPEG-Format mit verlustbehafteter Kompression in JPEG2000 (R 2) umgewandelt und anschließend in JPEG (R 3) zurückumgewandelt. Beide Formate JPEG2000 und JPEG stehen hier nur stellvertretend für künftige verlustbehaftete Bildkompressionsformate der Jahre 2056 und 2098.

Es fanden mehrere Versuchsreihen statt, die sich hinsichtlich der verwendeten Qualitätsparameter bei der Kompression unterschieden. Die Differenz wurde jedes Mal mit KOST-Simy festgehalten, wobei der Toleranzparameter «M» (für Medium) zur Anwendung kam. Mit diesem Toleranzparameter werden Fehler ausgeworfen, wenn mehr als 0,001 Prozent der Pixel sich um 5 Prozent oder mehr von ihrem früheren Farbwert unterscheiden. Abweichungen unterhalb dieser Schwellen werden nicht wiedergegeben.

Nr.	R0	P1	D1	R1	P2	D2	R2	S
1	3,7 MB	90	< 0,001 %	2,9 MB	100	< 0,001 %	5,4 MB	12,0 MB
2	3,7 MB	80	< 0,001 %	1,4 MB	90	< 0,001 %	1,6 MB	6,7 MB
3	3,7 MB	90	< 0,001 %	2,9 MB	90	< 0,001 %	1,5 MB	8,1 MB
4	3,7 MB	90	< 0,001 %	2,9 MB	80	0,04 %	0,8 MB	7,4 MB
5	3,7 MB	80	< 0,001 %	1,4 MB	80	0,05 %	0,8 MB	5,9 MB
6	3,7 MB	70	< 0,001 %	0,9 MB	80	0,07 %	0,8 MB	5,4 MB
7	3,7 MB	60	0,009 %	0,6 MB	100	0,01 %	4,3 MB	8,6 MB
8	3,7 MB	50	0,06 %	0,4 MB	50	1,13 %	0,4 MB	4,5 MB

Tabelle 1: Versuchsreihe über Veränderungen an Pixeln, die über einen farblichen Unterschied von 5% pro Pixel hinausgehen, nach Formatmigrationsprozessen unter verschiedenen Parametern. (R1: Ausgangsgröße, P1: Parameter JPEG2000, D1: Differenz zu R0, R1: Neue Dateigröße, P2: Parameter JPEG, D2: Differenz zu R0, R3: Neue Dateigröße, S: Summe Dateigrößen R0-R2)

Es lassen sich folgende Beobachtungen machen:

Wie die Versuchsreihen 1 bis 3 zeigen, ist es im Lauf zweier verlustbehafteter Formatmigrationen nicht nur möglich, die Farbwerte im vordefinierten Toleranzbereich zu halten, sondern gleichzeitig den Speicherbedarf der neuen Repräsentationen zu verringern.

Versuchsreihen 4 bis 6 zeigen, dass mit dem JPEG2000-Algorithmus bei der ersten Wandlung noch wesentlich geringere Dateigrößen erreichbar sind, ohne dass die von KOST-Simy gesteckte Grenze überschritten wird. Der zweite Migrationsschritt nach JPEG verfehlt dann aber die Schwelle des Erlaubten.

Versuchsreihen 3 und 7 zeigen, dass eine zu intensive Anwendung des einen Algorithmus unerwünschte Folgen hat, wenn das Einhalten der Grenze beim Folgealgorithmus das Ziel sein muss. Während in Versuchsreihe 7 der Speicherbedarf der R 2 mit 0,6 MB sehr niedrig wird, zieht der Speicherbedarf der R 3 mit 4,3 MB über den Ausgangsbedarf der R 1 hinaus an, weil ein hoher Qualitätsparameter erforderlich ist, um die Differenzgrenze einzuhalten. Es liegt nahe, dies auf die für jeden Algorithmus spezifische Form der Kompressionsartefakte zurückzuführen. Dass der Folgealgorithmus diese Muster mit seinen eigenen spezifischen Mustern in Einklang bringen muss, führt zu größeren Datenmengen. Deshalb erscheint es sinnvoll, das Maß an Artefakten in allen Formatmigrationen auf das Nötigste zu beschränken.

Insgesamt ist die Versuchsreihe 2 diejenige, die unter Einhaltung der gewünschten Qualitätsstandards eine Sicherung mit dem geringstmöglichen Speicherplatz (für die drei Repräsentationen zusammen) erlaubt.

Versuchsaufbau 2

Ein weiterer Versuch, der im Technischen Zentrum des Landesarchivs NRW in Münster durchgeführt wurde, bestand darin, die Grenzen des Schadens einer wiederholten Anwendung verlustbehafteter Kompressionsverfahren zu ermessen. Im Experiment wurde eine JPEG-Datei ohne zusätzliche inhaltliche Bearbeitung immer wieder mit dem gleichen höchsten Qualitätsfaktor neu verlustbehaftet komprimiert.

Anzahl Neukompressionen	Veränderung gegenüber R 1
10	1,419 %
100	3,865 %
1000	3,874 %

Tabelle 2: Versuchsreihe über Veränderungen an Pixeln, die über einen farblichen Unterschied von 5 % pro Pixel hinausgehen, nach mehrfacher JPEG-Kompression. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Hierbei zeigte sich, dass die Veränderungsrate, die sich aus häufig wiederholten JPEG-Kompressionen ergibt, nach vielen Wiederholungen immer geringer ausfällt. Das heißt, die ersten Wiederholungen führen zu stärkeren Veränderungen als die späteren, bis schließlich nur noch minimale Veränderungen feststellbar sind. Bemerkenswert an dem Ergebnis dieses Versuches ist vor allem, dass sich die bildlich nachweisbaren Veränderungen bei reinen Neukompressionen zumindest im gewählten Beispiel in recht engen Grenzen halten und nicht exponentiell «aufschaukeln».

Versuchsaufbau 3

Ebenfalls aus Münster stammt das Beispiel einer Rastergrafik, die sich von ihrem Pixelmuster kaum für eine verlustbehaftete Formatmigration eignet. Das Experiment zeigte, dass sich diese mangelnde Eignung auch in Kennzahlen nachweisen lässt. Es handelt sich um eine standardisierte Landkarte im Maßstab 1:25000 (DTK 25), die aus nur 18 verschiedenen, flächig angelegten Farbtönen besteht. Die Farbtiefe betrug 8 Bit, die Grafik war als TIFF-Datei mit einer proprietären RLE-Kompression von Apple verlustfrei komprimiert. Die JPEG-Konversion führte zu Veränderungen, die in Tabelle 3 dokumentiert sind.

Datei	TIFF (R 1)	JPEG (R 2)
Größe/Pixel	3200x3200	3200x3200
Kompression	RLE (Mac-Variante)	JPEG
Verwendete Farbwerte	18	36504
Größe	1,1 MB	5,7 MB
Veränderung zu R 1	-	9,282 %
Farbtiefe	8 Bit	24 Bit

Tabelle 3: Versuch der Formatmigration von TIFF nach JPEG, der sich anhand von automatisiert erhobenen Kennzahlen als Fehlschlag identifizieren lässt. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Dass eine JPEG-Umwandlung in diesem Fall kein erfolgversprechender Weg ist, zeigt sich an der erheblichen Veränderungsrate gegenüber dem Ausgangsbild, an der Vergrößerung der Datenmenge und an der Vervielfachung der Farbwerte. Interessant ist, dass bei dem hier vorliegenden standardisierten Bildtyp DTK 25 jede Vermehrung der effektiven Farbwerte über 18 hinaus einen Informationsverlust bedeuten würde, denn die Farbwerte haben im kartografischen Kontext jeweils eine exakte Bedeutung (z.B. Hellblau für Gewässer).

Versuchsaufbau 4

Der vierte Versuch, der im Staatsarchiv Ludwigsburg stattfand, beschäftigt sich mit dem Verhältnis zwischen rein digitalen Prozessen und den analog-digitalen Übergangsprozessen beim Scannen. Im Kollegenkreis waren die Verfasser darauf hingewiesen worden, dass die meisten Scanner nicht in der Lage sind, eine Vorlage zweimal in eine identische Bitfolge zu verwandeln.

Für den Versuch wurde eine aquarellierte Zeichnung im Folioformat zweimal unmittelbar hintereinander ohne Verrücken der Vorlage gescannt. Der Unterschied der beiden Scans betrug nach den oben (Versuch 1) genannten Maßstäben 1,388 % der Pixel. Bei anderen Scannermodellen, an denen der gleiche Versuch geplant war, war ein Vergleich der zwei erstellten Scans gar nicht möglich, weil diese Scanner jedes Mal eine leicht unterschiedliche Anzahl von Pixeln auswarfen.

Aus diesem Versuch ergibt sich, dass bei der Migration von einem Muster auf physischen Trägern in eine digitale Rastergrafik oft eine Variationsbreite vorliegt, die die Veränderungsrate bei verantwortungsvoll durchgeführten verlustbehafteten Formatmigrationen erheblich übersteigt. Mit anderen Worten: ein Scan hinterlässt unter Umständen viel mehr Qualitätsunsicherheiten als eine verlustbehaftete Formatmigration – nur weil dies bislang weniger bekannt ist, kann man es nicht ignorieren.

Abbildung 1: Ein daumennagelgroßer Ausschnitt aus dem Scan eines Aquarells.

Abbildung 2: Differenzbild zwischen Scan und JPEG-komprimierter Darstellungsform. Schwellenwerte wie in Versuchsaufbau 1, 3. Absatz.

Abbildung 3: Differenzbild zwischen Scan A und Scan B, ohne Verrücken der Vorlage. Schwellenwerte wie in Versuchsaufbau 1, 3. Absatz.

Versuchsaufbau 5

Ein oft vorgebrachtes Argument gegen komprimierte Dateiformate insgesamt zielt auf deren fehlende «Robustheit» ab. Gemeint ist damit die Toleranz eines Dateiformats gegen Bit-Rot, also: die ungeplante Veränderung von Bits durch technische Defekte. Tatsächlich lassen sich Unterschiede in der Robustheit empirisch nachweisen.

Im Technischen Zentrum des Landesarchivs NRW in Münster wurden hierzu jeweils 50 identische TIFF- und JPEG-Dateien einem künstlichen Bit-Rot unterzogen, indem einzelne Bits bzw. Bitfolgen verändert wurden. Danach wurden die nachweisbaren Bildinformationsverluste mit den bereits bekannten Mitteln gemessen.

Bit-Rot in %	Verlustquote in Pixeln TIFF (unkomprimiert)	Verlustquote in Pixeln JPEG
0,01 %	55,76 %	99,86%
0,001 %	10,46 %	97,43%
0,0002 %	2,81 %	92,79%

Tabelle 4: Auswirkungen von Bit-Rot auf die Bildqualität unkomprimierter und komprimierter Dateiformate. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Das Ergebnis des Versuchs ist eindeutig: Je dichter das Pixelmuster im Datenstrom komprimiert ist, desto schneller können auch kleine Veränderungen zu großen Problemen führen. So vertragen bei den hier in Frage stehenden Bildformaten unkomprimierte TIFF-Dateien im Schnitt eine deutlich höhere Anzahl fehlerbedingter Veränderungen als stark komprimierte JPEG-Dateien.[7]

Daraus zu folgern, dass unkomprimierte Dateiformate für den Langzeiterhalt per se besser seien als komprimierte, ist jedoch problematisch. Denn mit etwas Pech kann bei jedem Dateiformat die Veränderung bereits sehr weniger einzelner Bits zu einem kompletten Informationsverlust führen. Zudem basiert die Bewertung eines Dateiformats nach Robustheit auf der Annahme, Bit-Rot sei in einem Langzeitarchiv ein akzeptables oder zumindest ein unvermeidliches Phänomen. Dies entspricht weder dem fachlichen Diskussionsstand darüber, wie ein vertrauenswürdiges Langzeitarchiv einzurichten sei, noch der tatsächlichen Praxis. Vertrauenswürdige digitale Archive begegnen der Gefahr des Bit-Rot mit ausgereiften und bewährten Mechanismen der technischen Bitstream Preservation. Weswegen also sollte Robustheit als Qualitätskriterium für Langzeittauglichkeit aufrechterhalten werden? Der Archivinformatiker Gary McGath meinte dazu kürzlich in seinem Blog:

> *«Banning compression from archives in the hope of minimizing the damage from bit rot is a foolish preservation strategy.»*[8]

Gute Gründe für JPEG?

Die geschilderten Laborversuche, die sich problemlos um einige zusätzliche Szenarien erweitern ließen, haben den Fokus unserer bisherigen Überlegungen auf die Diskussion möglicher Risiken bestimmter Bildkompressionsverfahren gelegt, und zwar immer mit Blick auf die Gefahr des Informationsverlusts. Das Kriterium des

7 Vgl. hierzu auch: Heydegger, Volker: Just One Bit in a Million: On the Effects of Data Corruption in Files. In: Agosti, Maristella (u.a.) (Hg.): Research and Advanced Technology for Digital Libraries. ECDL 2009 Berlin / Heidelberg 2009, S. 315-326.

8 McGath, Gary: Bit-rot tolerance doesn't work, in: Mad File Format Science [Blog]: https://madfileformatscience.garymcgath.com/2016/10/18/bit-rot-tolerance/

(begrenzten) Informationsverlusts kann jedoch bei einer an den signifikanten Eigenschaften eines Objekts und der ökonomischen Leistungsfähigkeit eines Archivs ausgerichteten Bestandserhaltungsstrategie nur ein Entscheidungskriterium unter mehreren sein. In diesem abschließenden Kapitel sollen daher einige Aspekte angesprochen werden, die möglicherweise für die Archivierung verlustbehaftet komprimierter Dateien in einem digitalen Langzeitarchiv sprechen.

Konkret geschieht dies am Beispiel JPEG. Denn abgesehen davon, dass JPEG bereits ohnehin in den Archiven angekommen ist, gibt es eine ganze Reihe guter Gründe, JPEG als Langzeitformat in Betracht zu ziehen. So ist JPEG offen standardisiert[9] und äußerst weit verbreitet, was eine sehr hohe Restlebenszeit des Formats und eine geringe Verdrängungsgefahr erwarten lässt. Viewer, Konverter und Werkzeuge zur Qualitätssicherung sind in großer Zahl frei verfügbar – ein Vorteil, der JPEG insbesondere gegenüber JPEG2000 auszeichnet. Durch seine hohe Verbreitung wird die irgendwann anstehende Aufgabe der Formatmigration zudem große Teile der IT-Welt betreffen, zumindest, sofern diese ein Interesse am Erhalt älterer Datenbestände hat. Die Archive werden diese Aufgabe also technisch nicht alleine bewältigen müssen.

Last but not least bieten hochkomprimierte Dateiformate ökonomische Vorteile in allen Funktionsbereichen des digitalen Archivs. In der Archivwelt, die gerne (vermeintliche) fachliche Optimalstandards ohne eine kritische Überprüfung der Angemessenheit im Einzelfall einfordert, wird über den Einfluss ökonomischer Faktoren auf Entscheidungsprozesse nur ungern gesprochen. Dieses Ethos wird durch die deutsche Archivgesetzgebung zwar gestützt[10] – gleichwohl sind öffentliche Archive einer begrenzten Budgetierung unterworfen. Gerade in größeren Archiven werden diese Grenzen in der Begeisterung für einzelne Arbeitsvorhaben gerne ausgeblendet. Da sie aber trotzdem existieren und das Gesamtmaß der Gestaltungsmöglichkeiten bestimmen, besteht ohne eine reflektierte und verzahnte Steuerung des gesamten Arbeitsbereichs die Gefahr der «kalten Kassation». Anders ausgedrückt: In einer Welt begrenzter Ressourcen verhindert jede Entscheidung für eine Maßnahme die Durchführung einer anderen Maßnahme. Dieses Phänomen ist naturgemäß nicht zu verhindern, ließe sich jedoch durch eine flexible Prüfung fachlicher Angemessenheit unter Berücksichtigung der vorhandenen Ressourcen kontrollieren und strategisch steuern.

9 Die JPEG-Komprimierung ist normiert in ISO/IEC 10918.

10 So legt z.B. das Archivgesetz des Landes Nordrhein-Westfalen fest, dass über die Archivwürdigkeit angebotener Unterlagen «das zuständige Archiv unter Zugrundelegung fachlicher Kriterien» entscheidet (§ 2 Abs. 6 Satz 2 ArchivG NRW). Entsprechende Regelungen finden sich in den meisten anderen deutschen Archivgesetzen.

Bei der digitalen Archivierung könnte die Wahl eines Rasterbildformats, das den signifikanten Objekteigenschaften angemessen ist, eine wichtige Stellschraube gegen die «kalte Kassation» sein.

Hierzu ein konkretes Rechenbeispiel. In Nordrhein-Westfalen wird seit kurzem unter dem Dach des Lösungsverbundes «Digitales Archiv NRW» eine mandantenfähige digitale Archivierungslösung für Kommunen angeboten. Die Kosten für die Nutzung dieses Systems «DIPS.kommunal» belaufen sich für ein Archiv als Kunde voraussichtlich auf etwa € 20.000,- pro Jahr. In dieser Pauschalsumme enthalten sind 500 GB Netto-Datenspeicher; jedes weitere GB kostet € 3,12 pro Jahr.

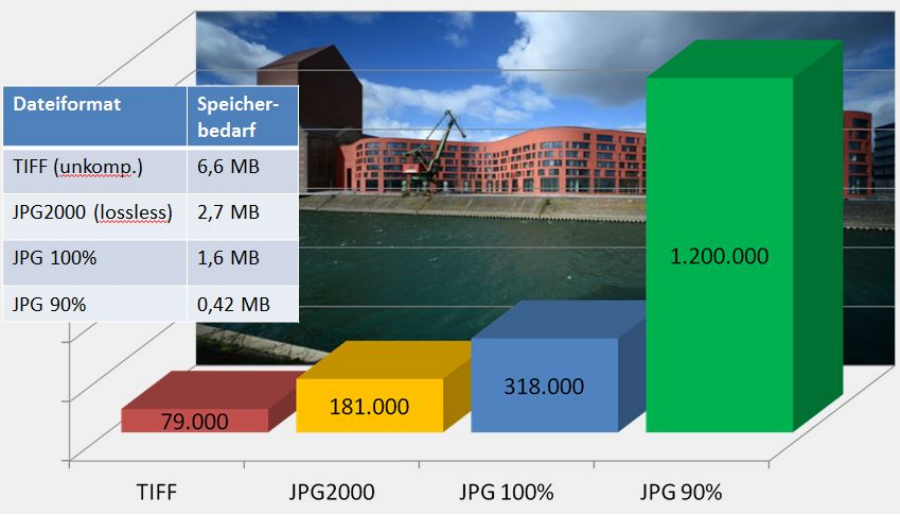

Abbildung 4: Beispielbild für unterschiedliche Speichervolumen

Das gezeigte Beispielbild ließe sich bei gleicher Auflösung als unkomprimierte TIFF-Datei mit einer Größe von 6,6 MB ca. 79.000 mal speichern, als verlustfreie JPEG2000-Datei mit einer Größe von 2,7 MB ca. 181.000 mal, als JPEG mit geringster Komprimierung und einer Größe von 1,6 MB rund 318.000 mal und als JPEG mit 90% Qualität und einer Größe von 0,42 MB etwa 1,2 Millionen mal. Die jährlichen Speicherkosten pro Bild betrügen (gerundet) € 0,25 (TIFF), € 0,11 (JPEG2000), € 0,06 (JPEG 100%), € 0,02 (JPEG 90%).

Der Einsatz verlustbehafteter Bildkompression kann somit in Archiven ohne unbegrenzte Ressourcen vor dem Totalverlust durch «kalte Kassation» schützen.

Fazit

Welche Erkenntnisse und Anregungen lassen sich nun aus den vorgestellten Überlegungen zum Umgang mit verlustbehafteten Bilddatenformaten festhalten?

Zunächst die Erkenntnis, dass ein geeignetes Instrumentarium für Vergleiche verschiedener Rasterbildformate zur Verfügung steht. ImageMagick und KOST-Simy ermöglichen sowohl eine vollautomatisierbare, in den Konvertierungsworkflow integrierbare Qualitätskontrolle für größere Datenmengen als auch händische Prüfungen mit Hilfe einer grafischen Benutzeroberfläche. Mit automatisiert erhobenen Kennzahlen lassen sich der Erfolgsgrad und der Umfang eingetretener Bildveränderungen einer verlustbehafteten Formatmigration bestimmen und bewerten. Auch lässt sich, basierend auf den Eigenschaften des Bilds, ein Optimum zwischen der Kennzahl Speicherbedarf und den Kennzahlen für Qualität ermitteln. Falsche Parameter können in diesem Prozess erkannt und korrigiert werden. Als allgemeine Regel könnte z.B. gelten: «Wenn ein Bild von Format A in Format B überführt wird, und sich mehr als 0,001 Prozent der Pixel verändern, sind Parameter oder Algorithmus ungeeignet.» Die Qualitätsparameter entsprechender Regeln sind bedarfsorientiert und nach fachlich wie ökonomischen Maßstäben zu bestimmen. Zum zweiten haben die Laborversuche verdeutlicht, dass die Entscheidung über den Einsatz verlustbehafteter Bilddatenformate am besten einzelfallabhängig zu treffen ist und von den angenommenen signifikanten Eigenschaften des digitalen Archivguts abhängig sein sollte.

Für die Weiterentwicklung des nestor-Leitfadens Bestandserhaltung wirft dies die Frage auf, ob die dort genannten, recht abstrakten Kriterien für die Auswahl von Bilddatenformaten mit den Erfüllungsgraden «Ja» / «Nein» /«für das menschliche Auge nicht erkennbar» in dieser Form praktisch ausreichend sein können. Als belastbarer hat sich in den Versuchen der Einsatz von Schwellenwerten erwiesen, die sich aus Kennzahlen zu Veränderungsgrad, Anzahl der Farbwerte und Datenmenge speisen. Es könnte daher sinnvoll sein, den im Leitfaden genannten Informationstyp «Bild» in zwei oder mehr Informationstypen aufzuspalten, zu denen dann passende Schwellenwerte definiert werden können. Nachdem schon Veronika Krauß 2016 in Potsdam wertvolle Anstöße in diese Richtung gegeben hat,[11] konnten auch die Verfasser dieses Beitrags nur einige weitere Schritte gehen.

Drittens hat sich in den Laborversuchen zumindest JPEG als ein Format erwiesen, das auch bei mehrfacher Kodierung und Dekodierung relativ widerständig

11 Krauß, Veronika; Bahrami, Arefeh: Ist das Bild noch das Bild? Authentizität digitaler Objekte unter Formattransformationen in Kooperation mit dem Thüringischen Hauptstaatsarchiv, Vortragsfolien von der 20. Tagung des Arbeitskreises Archivierung von Unterlagen aus digitalen Systemen am 1./2.3.2016 in Potsdam, online unter http://www.staatsarchiv.sg.ch/home/auds/20.html. Tagungsband erscheint demnächst.

gegen größere Veränderungen des Pixelbestandes ist. Insbesondere die mehrfache Neukodierung ohne willentliche Manipulation des Bildes führte zumindest im gewählten Beispiel nicht zu katastrophalen Pixelveränderungen. Umgekehrt formuliert: Die unabsichtliche Erzeugung von Artefakten, die mit dem menschlichen Auge erkennbar sind, ist schwieriger als erwartet. Ob die eingetretenen Veränderungen akzeptabel sind oder nicht, hängt freilich von den zuvor bestimmten signifikanten Eigenschaften des Informationsobjekts ab.

Viertens ist deutlich geworden, dass insbesondere im Umgang mit Digitalisaten die Bildqualität nur teilweise vom gewählten Bilddatenformat abhängig ist: So greift ein Scanprozess unter Umständen viel stärker und vor allem schwerer kalkulierbar in die Abbildungstreue einer Grafik ein als eine kontrolliert durchgeführte verlustbehaftete Formatmigration.

Last but not least hat sich gezeigt, dass der Einsatz eines verlustbehafteten Bilddatenformats wie JPEG auch einige gravierende Vorteile in der digitalen Langzeitarchivierung mit sich bringen kann. Wir glauben verstanden zu haben, dass die verlustbehaftete Kompression weniger Risiken birgt als bislang vermutet und einige Vorteile, insbesondere hinsichtlich der Langzeitverfügbarkeit des Formats und seiner ökonomischen Perspektiven, mit sich bringt. Gleichwohl bedarf es im Umgang mit verlustbehafteten Bilddatenformaten eines sehr sorgfältigen Risikomanagements, welches unbedingt auf genauen Kenntnissen der dem Format zu Grunde liegenden Technik basieren muss. Eine verlässliche Bestandserhaltung setzt voraus, dass man das Material, mit dem man es zu tun hat, gut kennt!

Wer genau hinschaut, stellt fest, dass verlustbehaftete Kompression nicht nur in den Reprowerkstätten der Archive schon längst zum Arbeitsalltag gehört. Langsam bildet sich ein intuitives Verständnis für ihre Möglichkeiten heraus, das uns zwar nicht vor kleinen Fehlern, aber vor großen Katastrophen bewahrt und uns vor dem Hintergrund hoher Speicherkosten ein Stück weit mehr Handlungsfähigkeit verschafft.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.7