Analyse JBIG2-Komprimierung
Fehlerhafte Xerox-Scans

1 Management Summary

1.1 Deutsch

Im August 2013 wurde ein beunruhigender Fehler beim Scannen von PDF-Dokumenten mit Xerox-Geräten festgestellt. Der Fehler besteht hauptsächlich darin, dass einzelne Ziffern durch andere Ziffern ersetzt werden, wie dies in Abbildung 1 ersichtlich ist. Diese falschen Ziffern sind pixelidentisch mit anderen Ziffern im Dokument.

Abbildung 1: Bildausschnitte vor und nach der JBIG2 Komprimierung[1]

Die unter anderem von Xerox verwendete verlustbehaftete JBIG2-Komprimierung speichert gleichwertige Symbole nur einmal ab und verwendet diese mehrfach im ganzen Dokument. Dieses Verfahren nennt sich "Pattern matching and substitution" (PMS). Mit dem PMS-Verfahren können signifikante Einsparungen in der Dateigrösse erreicht werden, ohne dass Kompressionsartefakte wie z.B. bei JPEG auftreten. Xerox hat bei der Implementierung zu stark auf die Dateigrösse fokussiert. Damit wurde zwar erreicht, dass die Dateigrösse sehr klein wurde, aber auch, dass verschiedene Zeichen als gleichwertig eingestuft und entsprechend ersetzt wurden.

Die KOST erstellte ergänzende Analysen zu diesem Thema und kommt zum Schluss, dass einzig die Anzahl Pixel pro Symbol (kleine Schriftgrösse und/oder geringe Scanauflösung) dafür massgebend ist, dass es bei den entsprechenden Xerox-Geräten zu diesem Substitution-Fehler kommt.

Nicht nur Xerox verwendet diese Komprimierung, sondern auch andere Hersteller wie z.B. Fujitsu. Es ist nicht nur ein Problem von Xerox, sondern vielmehr ein grundsätzliches Problem der JBIG2-Komprimierung.

In PDF/A-Dateien dürfen verlustbehaftete Komprimierungen wie zum Beispiel JBIG2 eingesetzt werden. Der grösste Teil der betroffenen PDF/A-Dateien stammt aus Digitalisierungen bei den abliefernden Stellen z.B. Posteingangsscanner. Bei einer Konvertierung von PDF zu PDF/A wird die JBIG2-Komprimierung und ggf. der Substitution-Fehler übernommen.

Da der Fehler irreversibel ist und nicht festgestellt werden kann, ob das PMS-Verfahren eingesetzt wurde oder nicht, empfiehlt die KOST, beim Erstellen von PDF-Dateien vorerst auf die Kompressionsart JBIG2 zu verzichten und die verschiedenen Quellen, insbesondere die Informatikstellen der abliefernden Stellen, zu sensibilisieren. Durch die Vermeidung von JBIG2 oder mindestens durch die Verwendung des neuen Xerox-Patches können die Anzahl fehlerhaften Scan-Dateien in Zukunft auf ein Minimum reduziert werden.

[1] Aus David Kriesel, "Xerox-Scankopierer verändern geschriebene Zahlen", http://www.dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning.

1.2 Français

En août 2013, une erreur préoccupante survenant lors de la copie numérisée de documents PDF avec des appareils Xerox a été constatée. L'erreur réside principalement dans la substitution de certains chiffres par d'autres comme le montre la figure 1. Ces chiffres erronés sont identiques à d'autres chiffres du document en termes de pixels.

Vorher		Nachher	
110.000	54,60	110.000	54,80
125.000	60,00	125.000	60,00
140.000	65,40	140.000	85,40
155.000	70,80	155.000	70,80
170.000	76,20	170.000	76,20

Figure 1: Fragments d'images avant et après la compression JBIG2[2]

La compression JBIG2 avec pertes, utilisée notamment par Xerox, ne sauvegarde qu'une seule fois les symboles de même valeur et les utilise plusieurs fois dans tout le document. Ce procédé est appelé *pattern matching and substitution* ou PMS (filtrage par motif et substitution). Le procédé PMS permet d'obtenir des compressions significatives de la taille des fichiers sans générer d'artéfacts de compression comme c'est le cas notamment avec JPEG. Xerox s'est trop focalisé sur la taille de fichier lors de l'implémentation, ce qui a permis d'obtenir une très petite taille de fichier, mais a également conduit à classer différents caractères comme étant de valeur identique et donc à les remplacer.

Le CECO a effectué des analyses complémentaires sur le sujet et est arrivé à la conclusion que seul le nombre de pixels par symbole (police de caractère de taille réduite et/ou basse résolution de numérisation) est déterminant pour générer cette erreur de substitution lors de l'utilisation des appareils Xerox concernés.

Xerox n'est pas le seul fabricant à utiliser cette compression, d'autres l'utilisent également, par exemple Fujitsu. Il ne s'agit pas uniquement d'un problème de Xerox, mais plutôt d'un problème fondamental de la compression JBIG2.

Dans les fichiers PDF/A les compressions avec pertes comme JBIG2 peuvent être utilisées. La plus grande partie des fichiers PDF/A concernés provient de numérisations réalisées auprès des services versants (par ex. scanners de courrier entrant). En cas de conversion de PDF en PDF/A, la compression JBIG2 et, le cas échéant, l'erreur de substitution sont reprises.

Comme l'erreur est irréversible et qu'il ne peut être détecté si le procédé PMS a été utilisé ou non, le CECO recommande d'éviter pour le moment le type de compression JBIG2 lors de l'établissement de fichiers PDF et de sensibiliser les différentes sources, en particulier les services informatiques des services versants. Éviter la compression JBIG2 ou au moins utiliser le nouveau correctif de Xerox permettra à l'avenir de réduire au minimum le nombre de fichiers de numérisation erronés.

[2] Tiré de David Kriesel, "Xerox-Scankopierer verändern geschriebene Zahlen", http://www.dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning

2 Das Problem

2.1 Einleitung

Am 1. August 2013 dokumentierte der deutsche Informatiker David Kriesel in seinem Blog einen beunruhigenden Fehler beim Scannen mit Xerox-Kopiersystemen[3]. Der Fehler besteht darin, dass in PDF-Dokumenten, die mit WorkCentre-Geräten von Xerox gescannt wurden, einzelne Ziffern oder ganze Blöcke gegenüber dem eingescannten Original verändert sind. Charakteristisch ist beispielsweise die Ersetzung einer "6" durch eine "8", wie in Abbildung 1 ersichtlich. Der Fehler konnte in der Folge auf diversen Xerox-Geräten[4] reproduziert werden.

Das Problem taucht bei TIFF-Scans nicht auf, sondern nur bei Scans in PDF. Eine genauere Analyse der falschen Ziffern ergab, dass diese pixelidentisch sind mit anderen Ziffern im gleichen Dokument. Die ursprünglichen Ziffern wurden also vom Scanverfahren passgenau ersetzt mit der Kopie einer anderen Ziffer im Dokument. Damit sind Pixelfehler beim Scannen und Probleme mit der OCR-Software als Fehlerquelle ausgeschlossen, und die Vermutung liegt nahe, dass das verwendete Komprimierungsverfahren die Ursache ist.

2.2 Analyse

2.2.1 Das Kompressionsverfahren JBIG2

Bei der Produktion der PDFs verwendet Xerox das Bildkompressionsverfahren JBIG2. Dieses teilt das Bild in drei sich überlappende Bereiche ein: Text, Grafiken und generische Regionen. Der entscheidende Gewinn beim Speicherplatz wird mit der Komprimierung des Textbereichs erzielt: "Eine Textregion besteht aus einer Anzahl von Symbolen, die auf einem Hintergrund platziert werden. Typischerweise entspricht ein Symbol einem Zeichen (z. B. Buchstaben), das in einem Text vorkommt. Die Symbole werden in einem Symbolwörterbuch gespeichert und können durch Angabe ihrer Indizes wiederverwendet werden. Die Speicherung im Wörterbuch erfolgt entweder als codierte Bitmap oder als Verfeinerung eines anderen Wörterbucheintrags, wobei lediglich die Differenz zur Vorlage gespeichert wird."[5]

Dieses Konzept wurde 1974 durch R. N. Ascher und G. Nagy entwickelt. Ihrer Publikation[6] entnehmen wir Abbildung 2, um den Einsatz eines Symbolwörterbuchs zu illustrieren:

---

[3] David Kriesel, "Xerox-Scankopierer verändern geschriebene Zahlen", http://www.dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning. Dieser Originalartikel wurde von Kriesel regelmässig aufdatiert und vermittelt einen Gesamtüberblick über das Problem und seine Analyse. Die Mitteilung wurde von der Fachpresse aufgenommen (z.B. "Xerox-Kopierer baut Zahlendreher in Scans", 03.08.2013 (Update 08.08.2013), http://winfuture.de/news,77252.html) und gelangte bald in die breitere Öffentlichkeit (z.B. "Lies, damned lies and scans", The Economist, 08.08.2013, http://www.economist.com/blogs/babbage/2013/08/perils-digital-technology).

[4] Laut Xerox (http://simplifywork.blogs.xerox.com/2013/08/07/update-on-scanning-issue-software-patches-to-come/ Eintrag vom 11.08.2013) sind folgende Geräte betroffenen:
ColorQube: 87XX, 89XX, 92XX, 93XX WorkCentrePro: 2XX BookMark: 40, 55
WorkCentre: 5030, 5050, 51XX, 56XX, 57XX, 58XX, 6400, 7220, 7225, 75XX, 76XX, 77XX, 78XX

[5] Wikipedia, JBIG2, http://de.wikipedia.org/wiki/JBIG2.

[6] R. N. Ascher, G. Nagy, "A means for achieving a high degree of compaction on scan-digitized printed text", in: IEEE Transactions on Computers, C-23:1174–1179, http://www.ecse.rpi.edu/~nagy/PDF_chrono/1974_Ascher_Nagy_IEEE_C74.pdf.

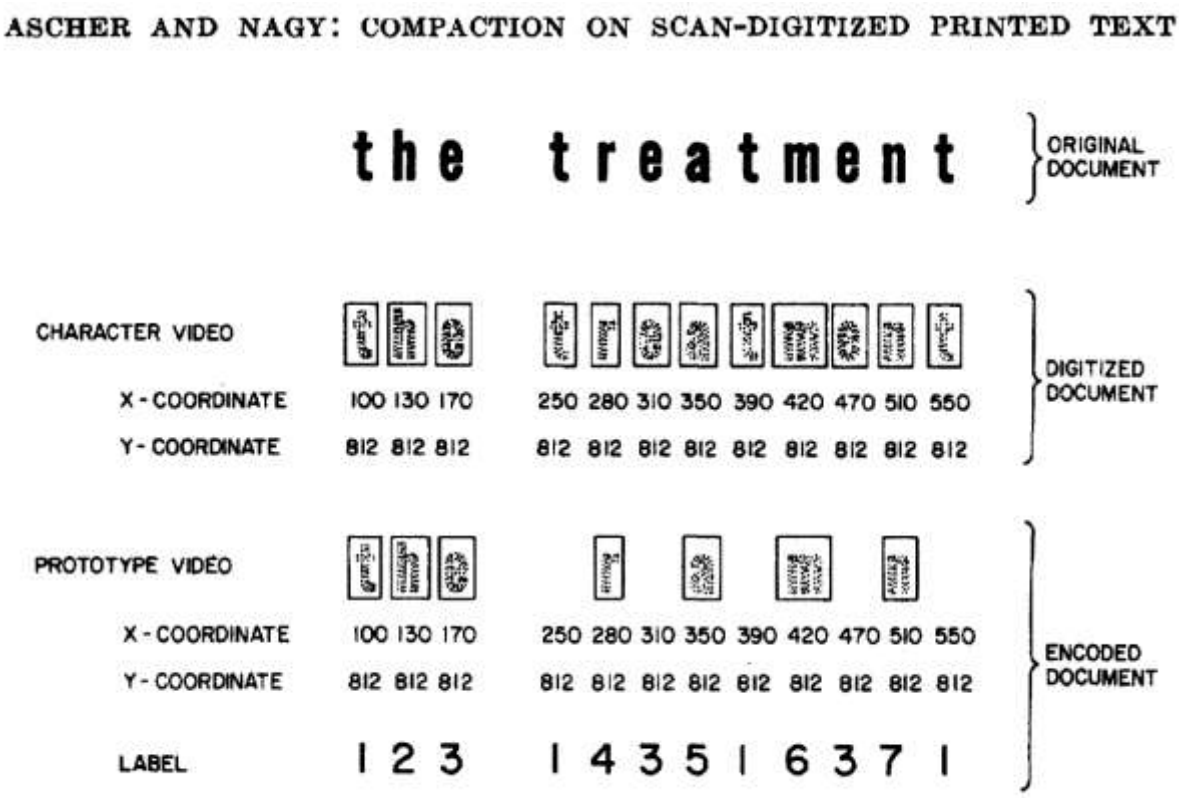


Fig. 1. Schematic diagram of video compaction method.

Abbildung 2: Illustration des Prinzips eines Symbolwörterbuchs

2.2.2 Verlustfrei oder verlustbehaftet

JBIG2 kann grundsätzlich auf zwei Arten implementiert werden, verlustfrei oder verlustbehaftet. Beide Arten verwenden die oben beschriebene Zeichenersetzung mittels Symbolwörterbuch ("Pattern matching and substitution"-Verfahren). Bei der verlustfreien Komprimierung müssen die Symbole zu 100 Prozent identisch sein, damit sie ersetzt werden. Bei der verlustbehafteten Komprimierung werden Symbole aus dem Symbolwörterbuch mehrfach verwendet, wenn sie "gleichwertig" sind. Mit dieser Methode kann JBIG2 vor allem bei Scans von Texten eine grosse Speicherplatzersparnis realisieren, was Xerox auch entsprechend vermarktet hat[7]. Zudem treten im Gegensatz zu anderen verlustbehaftenden Kompressionsverfahren wie zum Beispiel JPEG dabei keine typischen Kompressionsartefakte auf[8].Das Problem liegt jedoch offensichtlich bei der Definition von "gleichwertig". Wenn der Algorithmus dabei zu ungenau arbeitet, enthält das Resultat die eingangs beschriebenen falschen Zeichen (Substitutionsfehler).

[7] Cf. "Lies, damned lies and scans", oben Anm. 3
[8] Zur Illustration von JPEG-Kompressionsartefakten siehe Wikipedia, JPEG, http://de.wikipedia.org/wiki/JPEG.

2.2.3 ISO-Standard 14492

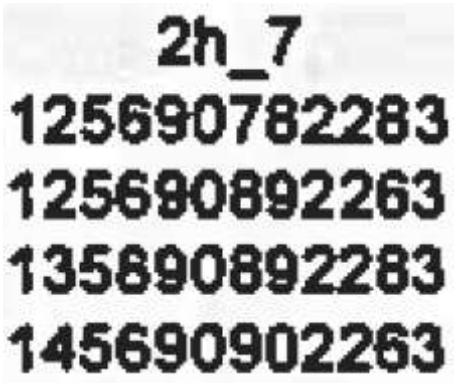
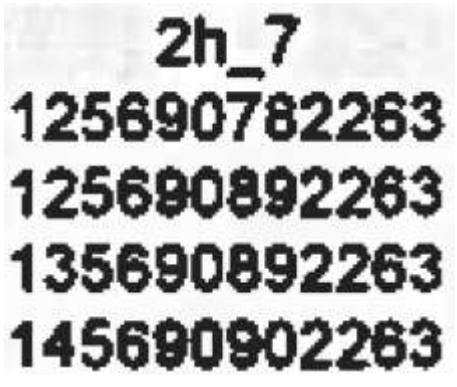
JBIG2 ist als ISO-Standard 14492 normiert[9]. Der Standard beschreibt jedoch lediglich, wie die Decodierung stattfindet, nicht wie der Encoder arbeiten muss. Für das vorliegende Problem hat dies zwei Konsequenzen:

- Es ist nicht im Standard festgehalten, sondern der Implementierung vorbehalten zu definieren, wann ein Symbol als gleichwertig gilt und bei der verlustbehafteten Komprimierung durch ein Vorgängerzeichen ersetzt werden darf.
- Im Resultat ist nicht in maschinenlesbarer Form ersichtlich, ob die verlustfreie oder verlustbehaftete Komprimierung angewendet wurde. Die beiden Arten können so beispielsweise von einem Validator nicht unterschieden werden. Allenfalls kann eine optische Analyse an Hand der Vielfalt der unterschiedlichen Zeichen Hinweise auf die Komprimierungsmethode geben; im Zweifelsfall ist aber von einer verlustbehafteten Komprimierung auszugehen.

2.3 Patch von Xerox

Ab dem 22.08.2013 hat Xerox Patches für die Firmware der betroffenen Geräte freigegeben und das Update empfohlen[10]. Die Patches deaktivieren die fehlerhafte Implementierung des Pattern Matching in allen Kompressionsmodi[11].

Tests der KOST mit der ersten Seite des Testdokumentes [A2] und der Einstellung "Archivieren" bestätigen, dass die fehlerhafte Implementierung des "Pattern matching and substitution"-Verfahren (PMS-Verfahren) offensichtlich nicht mehr verwendet wird oder die Implementierung von lossy auf lossless umgestellt wurde, weil fast keine Zeichen des Ergebnisses [B61] pixelidentisch sind und die Dateigrösse um 100% zugenommen hat. Nach dem Update des Xerox WorkCentre 7665 gibt es keine Substitution-Fehler mehr.

	Vor dem Update [B21]	Nach dem Update [B61]
Dateigrösse (nur Seite mit Arial 7)	55 KB	110 KB
Bildausschnitt 1		

[9] ISO/IEC 14492:2001: Information technology -- Lossy/lossless coding of bi-level images. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=22394

[10] Die Patches können unter http://www.xerox.com/scanpatch heruntergeladen werden; die entsprechenden Pressemeldungen finden sich unter http://realbusinessatxerox.blogs.xerox.com/2013/08/22/xerox_scanning_patch_available/.

[11] Siehe David Kriesel, "Xerox-Patch vorab getestet", 19.08.2013, http://www.dkriesel.com/blog/2013/0819_xerox_patch_tested.

3 Ergänzende Analysen der KOST

Die KOST-Geschäftsstelle hat die Analyse des Problems durch Xerox sowie durch interessierte Stellen weltweit verfolgt. Zusätzlich hat sie in eigenen, umfangreichen Untersuchungen einerseits den Fehler zu reproduzieren versucht, anderseits zusätzliche Fragestellungen formuliert und beantwortet. Sie erstellte dazu eigene Testdokumente und testete diese auf verschiedenen Geräten mit verschiedenen Einstellungen. Die Testvorlagen sowie die resultierenden Dokumente sind in Kapitel 5 Liste der Testdokumente aufgeführt und werden in der Folge mit ihren Nummern referenziert. Alle Dokumente sind zudem als Beilage zu dieser Studie in einer ZIP-Datei publiziert.

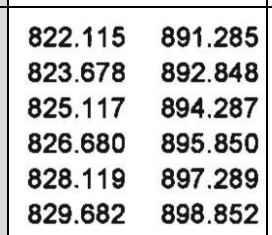
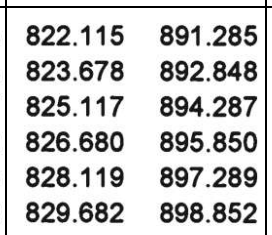
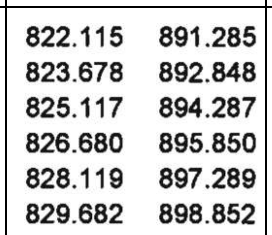
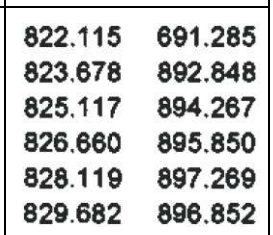
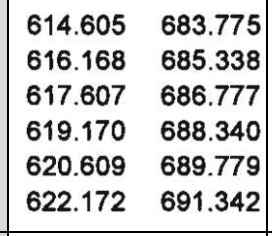
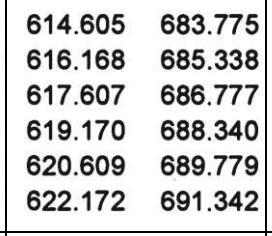
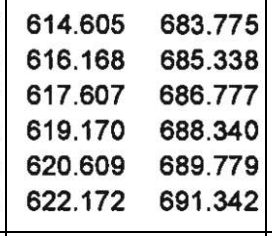
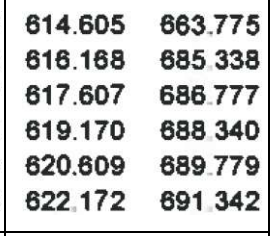
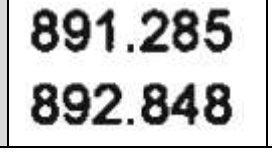
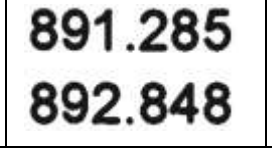
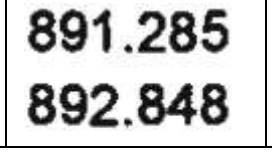
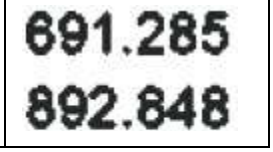
3.1 Analyse und Beurteilung Xerox WorkCentre 7665

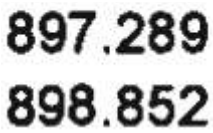
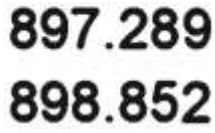
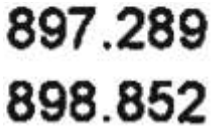
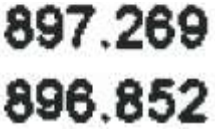
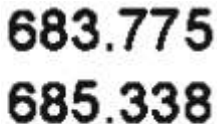
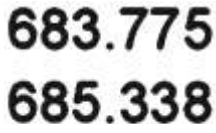
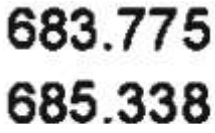
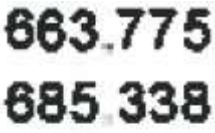
Die ersten Analysen wurden an einem Xerox WorkCentre 7665 durchgeführt, einem früh als betroffen identifizierten Gerätetyp. Sie dienten der Beantwortung von drei Fragen:

- Welches sind die Eigenschaften der verschiedenen Scaneinstellungen?
- Bei welchen Schriftgrössen tritt das Problem auf?
- Hat der Umfang des Textes eine Auswirkung auf den Umfang des Problems?

3.1.1 Analyse der verschiedenen Scaneinstellungen im Xerox WorkCentre 7665

Das Xerox WorkCentre 7665 bietet vier Standard-Scaneinstellungen an: "Standard", "Qualitativ hochwertig", "Freigeben und drucken" sowie "Archivieren". In der folgenden Tabelle sind für das Testdokument [A1] pro Einstellung die Dateigrösse, ein Bildausschnitt sowie eine Beschreibung des Ergebnisses ersichtlich.

Gewählte Einstellung	Standard [B11]	Qualitativ hochwertig [B12]	Freigeben und drucken [B13]	Archivieren [B14]
Dateigrösse	51 KB	178 KB	51 KB	37 KB
Scanauflösung	300dpi	600dpi	300dpi	200dpi
Bildausschnitt 1				
Bildausschnitt 2				
Detail 1a				

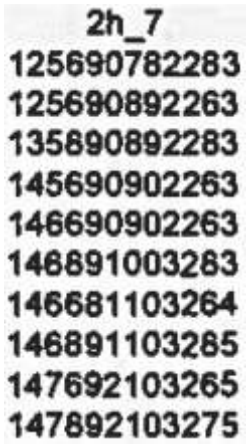
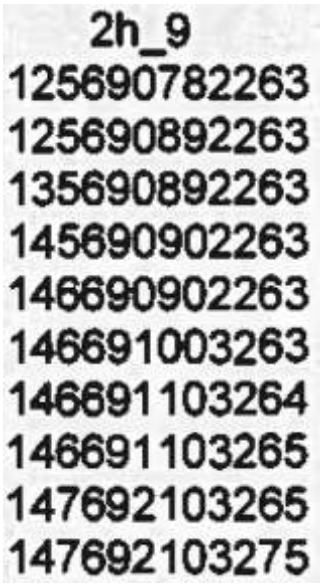
Gewählte Einstellung	Standard [B11]	Qualitativ hochwertig [B12]	Freigeben und drucken [B13]	Archivieren [B14]
Detail 1b				
Detail 2				
Ziffer "3" aus "683.775"	 Diese "3" besteht aus 14x22 Pixel.	 Diese "3" besteht aus 28x44 Pixel.	 Diese "3" besteht aus 13x22 Pixel.	 Diese "3" besteht aus 10x15 Pixel.
Ergebnis	Es sind keine Substitution-Fehler vorhanden. Dennoch wurde offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt, da zum Beispiel die einzelnen Ziffern aus "683.775" mehrfach vorkommen [C11].	Es sind keine Substitution-Fehler vorhanden. Dennoch wurde offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt, da zum Beispiel die einzelnen Ziffern aus "683.775" mehrfach vorkommen [C12].	Es sind keine Substitution-Fehler vorhanden. Dennoch wurde offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt, da zum Beispiel die einzelnen Ziffern aus "683.775 mehrfach vorkommen [C13].	Das verlustbehaftete PMS-Verfahren wurde eingesetzt, da mehrere Substitution-Fehler vorhanden sind. Auftretende Fehler nach ihrer Häufigkeit: "6" statt "8" "8" statt "9" "8" statt "6" [C14].

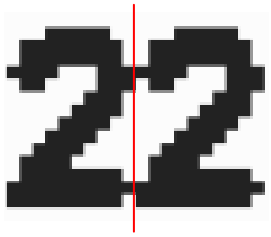
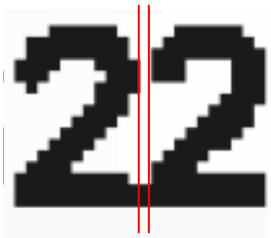
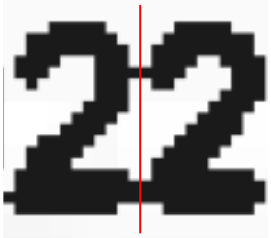
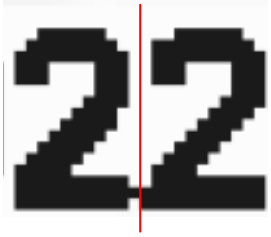
Das verlustbehaftete "Pattern matching and substitution"-Verfahren (PMS-Verfahren) wurde offensichtlich bei allen Einstellungen verwendet. In unserem Test traten jedoch nur mit der Einstellung "Archivieren" Substitution-Fehler auf. Der Unterschied zwischen den verschiedenen Einstellungen liegt in der Grösse eines Pixels, sprich der Scanauflösung. Die Implementierung des JBIG2-Verfahrens ist offensichtlich die gleiche für alle Einstellungen.

3.1.2 Analyse der verschiedenen Schriftgrössen im Xerox WorkCentre 7665

Die betroffenen Passagen in den Dokumenten von David Kriesel sind in Arial, Schriftgrösse 7 gesetzt. Es liegt nahe, dass in erster Linie kleine Schriftgrössen kritisch sind. Die KOST hat deshalb untersucht, ab welcher Schriftgrösse das Problem nicht mehr auftritt.

In der folgenden Tabelle sind für das Testdokument [A2] die Ergebnisse der verschieden Schriftgrössen in Arial ersichtlich. Gescannt wurde mit der Einstellung "Archivieren".

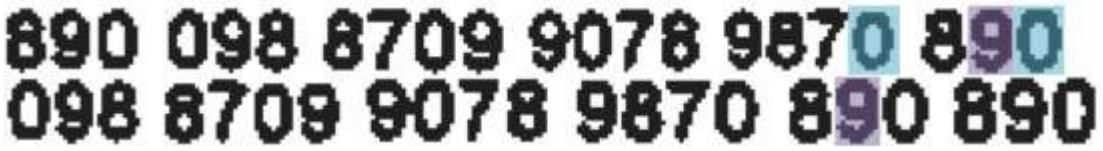
Schrift	Arial 7pt [B21]	Arial 8pt [B21]	Arial 9pt [B21]
Bildausschnitt 2h			
Ersten zwei Zeichen aus Block 2h			
Detail 1	 Eine "3" besteht aus 10x15 Pixel.	 Diese "3" besteht aus 11x17 Pixel.	 Diese "3" besteht aus 12x19 Pixel.

Schrift	Arial 7pt [B21]	Arial 8pt [B21]	Arial 9pt [B21]	
Analyse der verketteten 22	 Trennt man diese "22", sind die beiden "2" pixelidentisch und diese "2" kommt auch an anderen Stellen vor, z.B. zwei Zeilen weiter unten.	 Egal wie man diese "22" trennt, eine pixelidentische "2" konnte im ganzen zweiten Zeilenblock nicht gefunden werden.  Egal wie man diese "22" trennt, eine pixelidentische "2" konnte im ganzen zweiten Zeilenblock nicht gefunden werden – jedoch eine pixelidentische "22" zwei Zeilen weiter unten.  Egal wie man diese "22" trennt, eine pixelidentische "2" konnte im ganzen zweiten Zeilenblock nicht gefunden werden - jedoch eine pixelidentische "22" zwei Zeilen weiter unten.		-
Ergebnis	Detail 1: • Alle Ziffern "3" sind pixelidentisch. • Die beiden Ziffern "8", welche anstelle der "6" eingefügt wurden, sind auch pixelidentisch. • Einige Ziffern "2" sind ebenfalls pixelidentisch. Es traten wieder Substitution-Fehler auf.	Detail 1: • Die zweite "22" ist pixelidentisch mit der vierten. • Die verbundenen 22 sind mit hoher Wahrscheinlichkeit als ein einzelnes Zeichen abgespeichert. Es sind keine Substitution-Fehler vorhanden. Es wurde dennoch offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt, da auch bei Schriftgrösse 8pt bis zu zwei verbundene Ziffern pixelidentisch sind.	Detail 1: • Alle Ziffern "3" sind pixelidentisch. • Einige Ziffern "2" und "6" sind ebenfalls pixelidentisch. Es sind keine Substitution-Fehler vorhanden. Es wurde dennoch offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt, da auch bei Schriftgrösse 9pt ganze Ziffern pixelidentisch sind.	

Ein Zeichen wird im PMS-Verfahren als ganzes Symbol erkannt; grössere Schriftzeichen werden nicht geteilt. Bei Schriftgrösse Arial 8pt und grösser taucht der Substitution-Fehler nicht mehr auf. Dies hat jedoch offensichtlich nichts mit der Symbolgrösse des PMS-Verfahrens zu tun, sondern eher mit der Anzahl Pixel pro Zeichen respektive PMS-Symbols.

3.1.3 Analyse mit nur wenig identischen Zeichen im Xerox WorkCentre 7665

Eine weitere Hypothese ist, dass der Substitution-Fehler umso häufiger auftaucht, je mehr Zeichen auf einer gescannten Seite vorhanden sind. Im Umkehrschluss lautet diese Hypothese, dass unter einer bestimmten Anzahl Zeichen kein Substitution-Fehler entsteht. Dazu wurden Dokumente [A3] mit wenig Text in Arial 7pt mit der Einstellung "Archivieren" gescannt. In der folgenden Tabelle sind die Ergebnisse [B31]-[B33] ersichtlich.

<table>
<tr><th>Schrift</th><th>Originalausschnitt</th><th>Scanausschnitt</th></tr>
<tr><td rowspan="2">Je 1 Zeile [B31]</td><td>0987654321 2143658709
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM</td><td>0987654321 2143658709
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM</td></tr>
<tr><td colspan="2">Kein Substitution-Fehler. Mit einer Ausnahme (9) sind alle je sechs "8", "9" und "0" einmalig.
  </td></tr>
<tr><td rowspan="2">Je 2 Zeilen [B32]</td><td>tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
0987654321 2143658709
2143658709 9078563412
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM</td><td>tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
0987654321 2143658709
2143658709 9078563412
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM</td></tr>
<tr><td colspan="2">Kein Substitution-Fehler. Mit zwei Ausnahmen ("0", "9") sind alle "8", "9" und "0" einmalig.
  </td></tr>
<tr><td rowspan="2">Je 3 Zeilen [B33]</td><td>tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
0987654321 2143658709
2143658709 9078563412
9078563412 3216549870
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM</td><td>tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
0987654321 2143658708
2143658709 9078563412
9078563412 3216549870
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
tuvwxyz abcdefghijklmnopqrstuvwxyz
OPQRSTUVWXYZ ABCDEFGHIJKLM
  </td></tr>
<tr><td colspan="2">Es trat ein Substitution-Fehler auf: Bei der Zahlenreihe "2143658709" rechts wurde aus der Ziffer "9" eine "8". Diese "8" ist pixelidentisch mit der eigentlichen "8" aus dieser Zahlenreihe, welche insgesamt elfmal auftaucht. Neben dieser Ziffer "8" existieren vier weitere Varianten der Ziffer "8", welche viermal, zweimal oder einmal vorkommen (insgesamt 19 [statt korrekt 18] Vorkommen der Ziffer "8").</td></tr>
</table>

Schrift	Originalausschnitt	Scanausschnitt
Je 4 Zeilen [B34]		
	Es traten mehrere Substitution-Fehler auf. • Bei der Zahlenreihe "2143658709" rechts wurde aus der Ziffer "6" eine "8" und aus der Ziffer "8" eine "6". • Bei der Zahlenreihe "0987654321" links wurde aus der Ziffer "6" eine "8", ebenso bei der Zahlenreihe "3216549870" links. • Beim kleinen Alphabet wurde einmal ein kleines "i" durch ein grosses "I" ersetzt. • Ausserhalb des oben gezeigten Bildausschnittes wurde aus einem "t" eine "1". Bei allen Fehlern handelt es sich um Substitution-Fehler, da diese Zeichen pixelidentisch mit anderen Zeichen sind.	

Wie erwartet führt die zunehmende Anzahl gleicher Zeichen zu einer vermehrten Anwendung des verlustbehafteten "Pattern matching and substitution"-Verfahren; entsprechend ist auch die Wahrscheinlichkeit für einen Substitution-Fehler höher.

3.1.4 Beurteilung

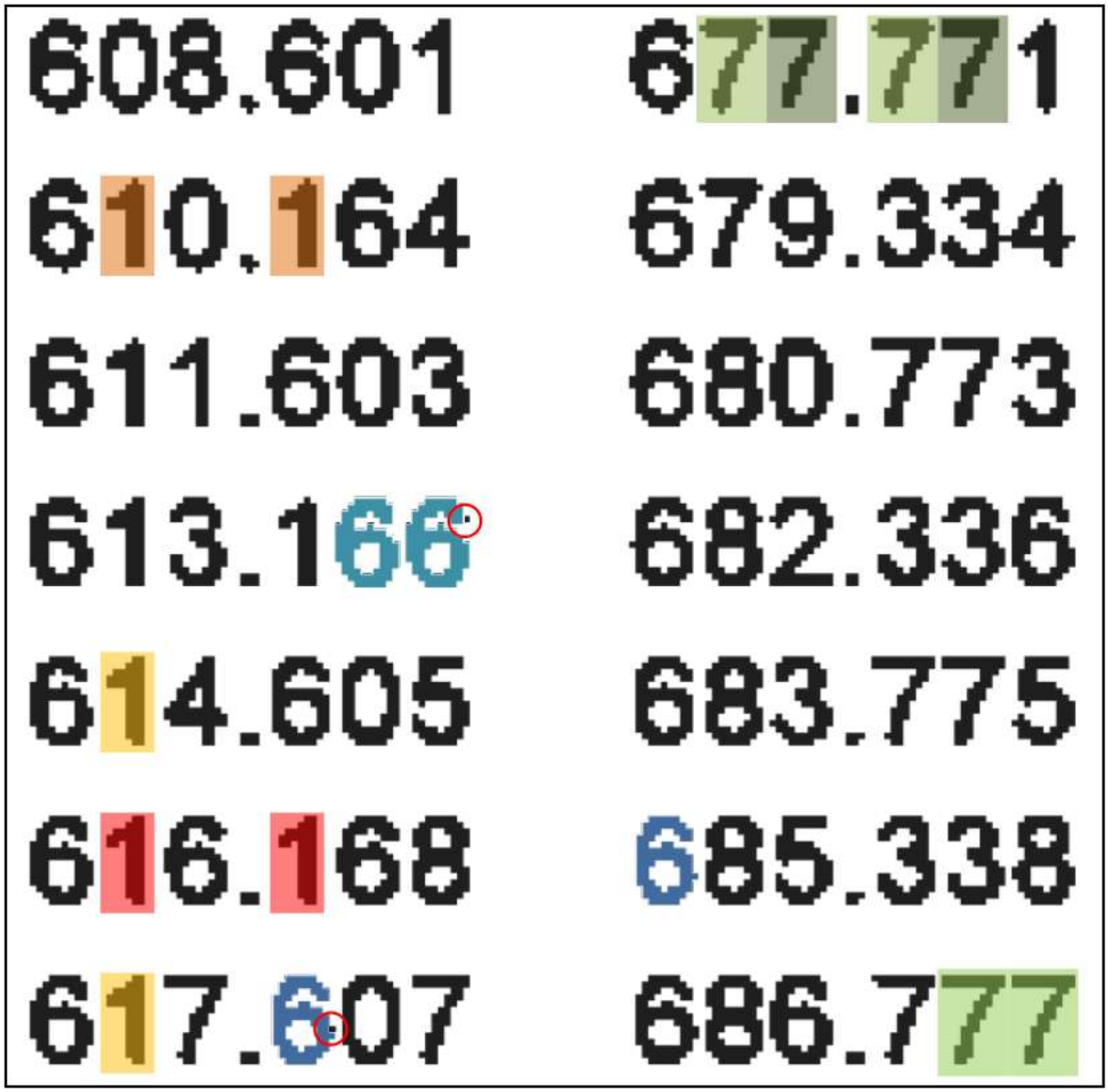
Xerox informierte nie über Details der Implementierung der JBIG2-Komprimierung in ihren WorkCentre-Geräten. Die Testresultate der KOST lassen den Schluss zu, dass nicht die Grösse des Zeichens alleine ausschlaggebend ist, sondern vielmehr die Anzahl Pixel, aus denen ein solches Zeichen besteht. Es ist wahrscheinlich, dass das Symbol als gleich angesehen wird, wenn die Anzahl unterschiedlicher Pixel nicht zu gross ist.

3.2 Analyse & Beurteilung anderer Scanner mit JBIG2

Als Implementationsvergleich wurden zwei weitere Scanner getestet, welche ebenfalls die JBIG2-Komprimierung benutzen.

3.2.1 Analyse der Implementierung im Xerox WorkCentre 7346

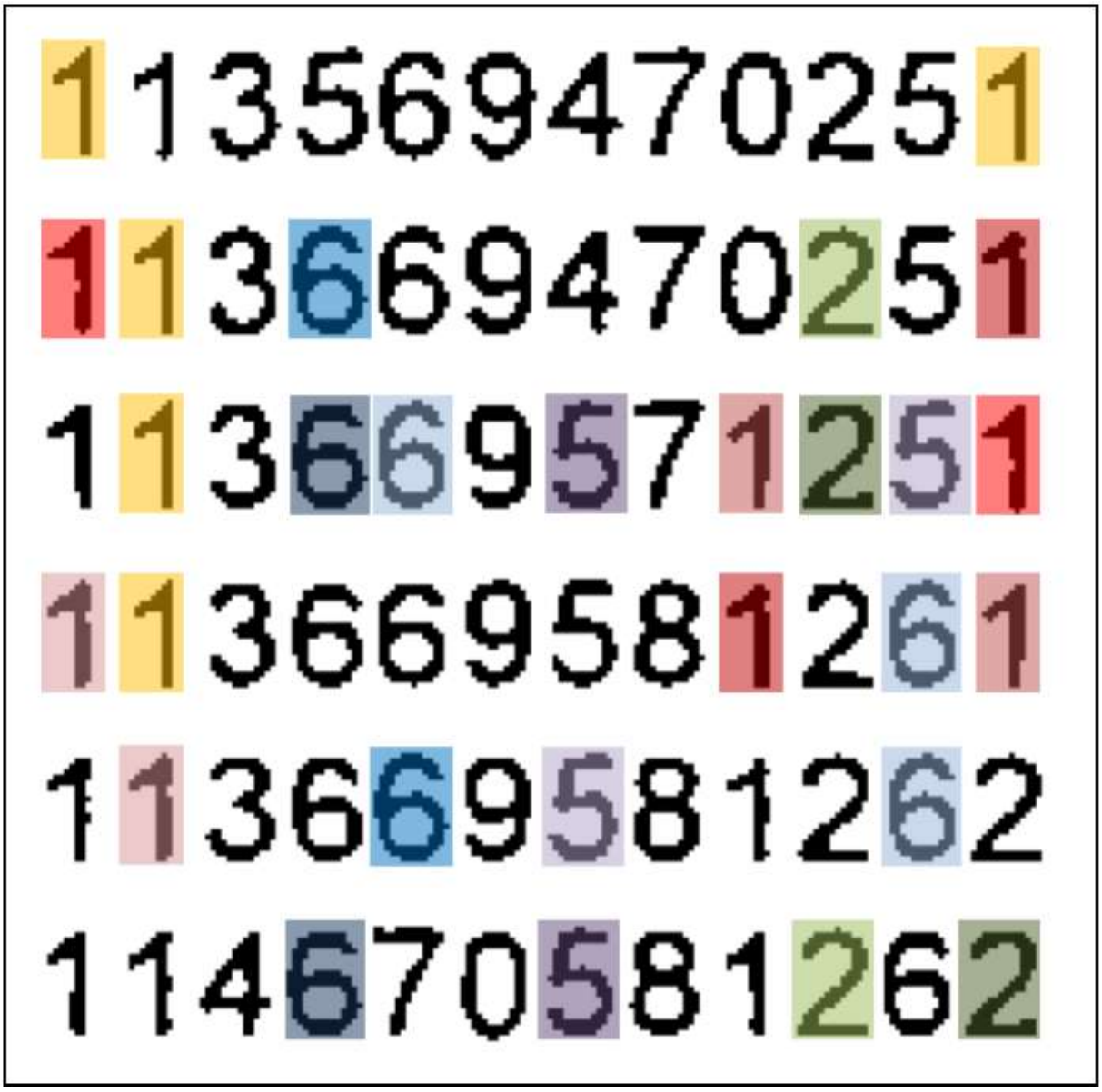
Für die folgende Analyse wurde die Vorlage [A1] in ein PDF gescannt mit JBIG2-Komprimierung (arithmetisch) und der Qualität Normal (niedrigste Qualität) [B15]. Daraus wurden ein Bereich mit vielen Ziffern "6" ausgewählt und die häufigsten Ziffern daraus ("6": 25mal, "7": 14mal sowie "1": 12mal) optisch miteinander verglichen.



In diesem Ausschnitt gibt es jeweils drei pixelidentische "7"-er und "1"-er Paare. Die 25 Instanzen der Ziffer "6" unterscheiden sich immer mindestens um einen Pixel. Es ist gut möglich dass die pixelidentischen "7" und "1" auch im Original-Scan (d.h. bevor die JBIG2-Komprimierung vorgenommen wurde) "pixelidentisch" gewesen sind, da diese Zeichen einfacher aufgebaut sind als die restlichen. Die Resultate, insbesondere die Vielfalt an Zeichen, legen die Vermutung nahe, dass dieses JBIG2-Implementierung nicht das verlustbehaftete "Pattern matching and substitution"-Verfahren einsetzt und entsprechend verlustfrei sein könnte. Zudem ist dieser Xerox-Scanner nicht auf der Liste der betroffenen Geräte.

3.2.2 Analyse der Implementierung im Fujitsu fi-6130

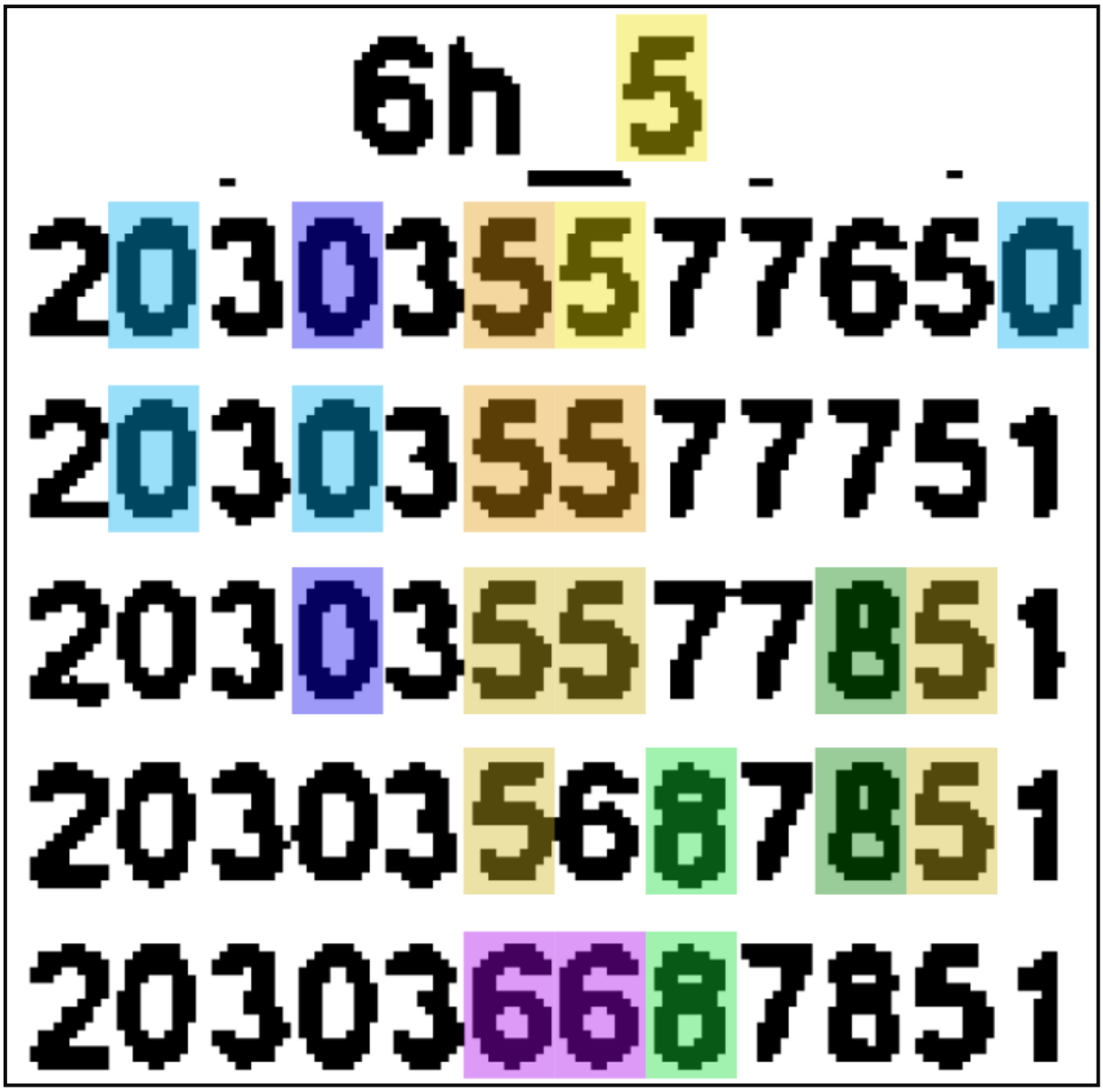
Für die nachfolgende Analyse wurde die Vorlage [A4] mit dem Fujitsu-Scanner fi-6130 in ein PDF gescannt. Dabei wurden keine Einstellungen verändert [B41].



Die optische Analyse des Bildausschnittes (zweitletzte Spalte, Zeilen 1 bis 6) zeigt, dass offensichtlich das verlustbehaftete PMS-Verfahren eingesetzt wurde, da nicht nur einfache Zeichen ("1") sondern auch komplexere Zeichen ("2", "5" und "6") pixelidentisch sind und sich keine Zeichen finden, welche sich nur durch ein einzelnes Pixel unterscheiden.

Mit den unveränderten Einstellungen konnte kein Substitution-Fehler entdeckt werden. Zudem besteht die "3" aus 12x20 Pixel, welches vergleichbar ist wie der Scan [B21] (Xerox WorkCentre 7665, Archivieren, Arial 9pt).

Da die Einstellungen nicht verändert werden dürfen, wurde für eine weitere Analyse eine Vorlage mit Arial 5pt [A5] erstellt. Diese wurde mit dem Fujitsu-Scanner fi-6130 in ein PDF gescannt [B42] und die Ziffern "0", "5", "6" und "8" optisch verglichen.

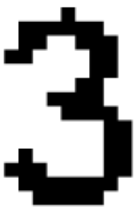


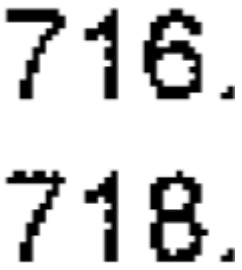
Mit den unveränderten Einstellungen konnte kein Substitution-Fehler entdeckt werden. Zudem besteht die "3" jetzt aus 9x16 Pixel, welches vergleichbar ist wie der Scan [B21] (Xerox WorkCentre 7665, Archivieren, Arial 7pt). Es ist dennoch nicht auszuschliessen, dass mit einer veränderten Einstellung auch dieser Scanner einen Substitution-Fehler produzieren würde. Diese Aussage stützt sich einerseits auf die grosse Anzahl gefundener pixelidentischer Zeichen im Bildausschnitt, anderseits auf etliche Hinweise in Publikationen, welche auf diese Problematik aufmerksam machen[12].

[12] Zum Beispiel wird auch im letzten Entwurf der JBIG2-Spezifikation im Kapitel 0.2.1 Symbol coding darauf hingewiesen (http://www.jpeg.org/public/fcd14492.pdf).

3.3 Analyse der Konvertierung TIFF zu PDF

Bei dieser Analyse wurde zuerst mit dem Xerox-WorkCentre 7346 und mit IrfanView ein schwarz/weiss TIFF mit der "CCITT G4"-Komprimierung der Testdatei A1 erstellt. Dieses TIFF dient als Testdatei [A6] für die Konvertierung in ein PDF mit Adobe Acrobat X Pro (Version 10.1.7). Dabei wurden verschiedene Komprimierungen für monochrome TIFF-Dateien getestet und verglichen, nämlich JBIG2 (verlustfrei), JBIG2 (verlustbehaftet) sowie CCITT G4.

Gewählte Einstellung	TIFF [A6] CCITT G4	PDF [B51] JBIG2 (lossless)	PDF [B52] JBIG2 (lossy)	PDF [B53] CCITT G4
Dateigrösse	79 KB	52 KB	30 KB	83 KB
Bildaus-schnitt 1	822.115 891.285 823.678 892.848 825.117 894.287 826.680 895.850 828.119 897.289 829.682 898.852	822.115 891.285 823.678 892.848 825.117 894.287 826.680 895.850 828.119 897.289 829.682 898.852	822.115 891.285 823.678 892.848 825.117 894.287 826.680 895.850 828.119 897.289 829.682 898.852	822.115 891.285 823.678 892.848 825.117 894.287 826.680 895.850 828.119 897.289 829.682 898.852
Bildaus-schnitt 2	614.605 683.775 616.168 685.338 617.607 686.777 619.170 688.340 620.609 689.779 622.172 691.342	614.605 683.775 616.168 685.338 617.607 686.777 619.170 688.340 620.609 689.779 622.172 691.342	614.605 683.775 616.168 685.338 617.607 686.777 619.170 688.340 620.609 689.779 622.172 691.342	614.605 683.775 616.168 685.338 617.607 686.777 619.170 688.340 620.609 689.779 622.172 691.342
Detail 1a	891.285 892.848	891.285 892.848	891.285 892.848	891.285 892.848
Detail 1b	897.289 898.852	897.289 898.852	897.289 898.852	897.289 898.852
Detail 2	683.775 685.338	683.775 685.338	683.775 685.338	683.775 685.338
Ziffer "3" aus 683.775				

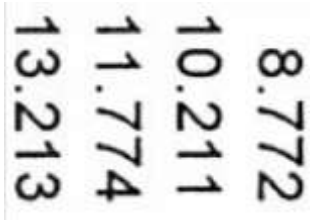
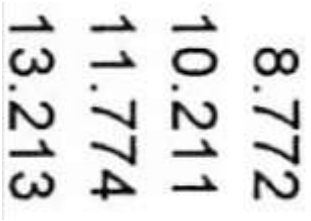
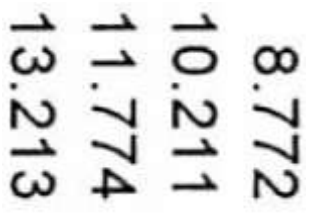
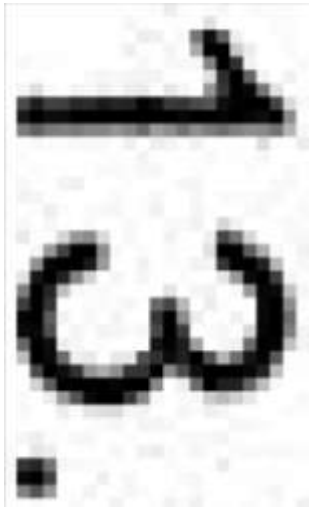
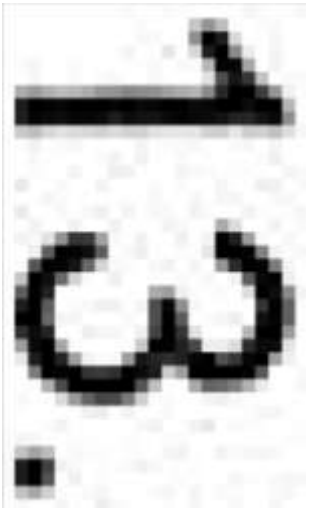
Gewählte Einstellung	TIFF [A6] CCITT G4	PDF [B51] JBIG2 (lossless)	PDF [B52] JBIG2 (lossy)	PDF [B53] CCITT G4
Ergebnis	Bereits im Ausgangsdokument im TIF-Format können Zeichen pixelidentisch sein: 	Die einzelnen Zeichen sind pixelidentisch mit dem TIFF. Das verlustbehaftete PMS-Verfahren wurde nicht eingesetzt.	Die einzelnen Zeichen sind nicht pixelidentisch mit dem TIFF. Das verlustbehaftete PMS-Verfahren wurde eingesetzt. Es wurden Zeichenersetzungen vorgenommen, jedoch tauchen keine Substitution-Fehler auf. Durch die vorgenommenen Zeichenersetzungen konnte die Dateigrösse nochmals wesentlich verkleinert werden.	Die einzelnen Zeichen sind pixelidentisch mit dem TIFF. Die Dateigrösse ist grösser als diejenige von [B51] (JBIG2 lossless). Zudem ist sie auch grösser als das Original [A6], welche die gleiche Komprimierung verwendet aber keinen PDF-Header hat.

Dieser Test zeigt abschliessend sehr schön den Unterschied zwischen verlustfreier und verlustbehafteter Komprimierung: Bei der verlustbehafteten Komprimierung sind Zeichen nicht mehr pixelidentisch, d.h. einige Pixelinfomationen wurden verändert. Im Gegenzug konnte die Dateigrösse stark reduziert werden.

3.4 Scanqualität und Pixelidentität

In den vorgängigen Untersuchungen wurde stillschweigend davon ausgegangen, dass ein Scanprozess ein bestimmtes Pixelbild eines analogen Originals liefert. Dem ist genaugenommen nicht so; und das hat einen grossen Einfluss auf die Beurteilung von verlustfreien bzw. verlustbehafteten Komprimierungen.

Bei dieser Analyse wurde die Testdatei A1mit dem Scanner Xerox WorkCentre 7346 mehrfach hintereinander eingescannt, ohne dass die Vorlage dazwischen bewegt wurde. Dabei soll aufgezeigt werden, das der wiederholte Scanprozess selber bereits nicht pixelidentische Ergebnisse liefert, wie es theoretisch[13] und praktisch[14] schon publiziert wurde.

	1. Scan [B71]	2. Scan [B72]	3. Scan [B73]
Bildausschnitt 1			
Detail 13. aus 13.213			

Technischer Hintergrund ist einerseits das Rauschen in den Aufnahmesensoren und bei der analog-digital Wandlung und andererseits eine gewisse Ungenauigkeit in der Scannermechanik.

Wie diese Analyse zeigt, sind geringfügige Pixelabweichungen dem Scanner bzw, dem Scanprozess zuzuordnen und müssen in Kauf genommen werden.

[13] William Palmer, Peter May, Peter Cliff, "An Analysis of Contemporary JPEG2000 Codecs for Image Format Migration", in iPRES2013, http://purl.pt/24107/1/iPres2013_PDF/An%20Analysis%20of%20Contemporary%20JPEG2000%20Codecs%20for%20Image%20Format%20Migration.pdf.

[14] Ralf Stockmann, "Bild-Datenkomprimierung aus Sicht der Nutzer", in KOST-Kolloquium "Datenkomprimierung bei Bild, Audio und Video", 2011, http://kost-ceco.ch/cms/index.php?compression_de

4 Beurteilung und Empfehlung

4.1 Beurteilung

Grundsätzlich ist gegen die verlustbehaftete Komprimierung mit geringfügigen Pixelabweichungen beim Scannen nichts einzuwenden, da der Scanprozess selber geringfügigen Pixelabweichungen mit sich bringt und, wie die oben erwähnten Studien zeigen, die Pixelabweichung zwischen zwei folgenden Scannvorgängen grösser ist als die Pixelabweichung zwischen dem ersten Scan und dem daraus verlustbehaftet komprimierten PDF. Inakzeptabel ist jedoch ein fehlerhaft implementiertes "Pattern matching and substitution"-Verfahren, das zum Vertauschen von Zeichen führen kann.

Die Untersuchungen von David Kriesel und anderen sowie die Analysen der KOST zeigen klar auf, dass die verlustbehaftete JBIG2-Komprimierung das Risiko fehlerhafter Zeichen in den gescannten Dokumenten mit sich bringt. Da aus einer PDF-Datei nicht hervorgeht, ob die verlustbehaftete oder die verlustfreie JBIG2-Komprimierung verwendet wurde, unterliegt grundsätzlich jede mittels JBIG2 erstellte Datei einem Generalverdacht. Von Revisionssicherheit kann bei solchen Dateien deshalb nicht mehr gesprochen werden. Einzig eine menschliche, inhaltliche Analyse des Dokuments kann gegebenenfalls (d.h. wenn keine kleinen Schriftgrössen verwendet wurden) eine hinreichende Sicherheit für Korrektheit ergeben. Der Fehler ist irreversibel, das heisst, dass auch eine Konvertierung diesen Fehler nicht beheben kann.

In PDF/A-Dateien darf verlustbehaftete Komprimierung wie zum Beispiel JBIG2 eingesetzt werden. Die ISO-Norm 19005 hält einleitend ausdrücklich fest, PDF/A stelle nicht unbedingt sicher, dass das digitale Erscheinungsbild dem Original entspricht, und empfiehlt betroffenen Organisationen, diese Risiken durch Policies und Best Practices abzusichern[15].

4.2 Empfehlung

Aus der Beurteilung lässt sich folgende Empfehlung ableiten:

Aus Sicherheitsgründen sollte beim Erstellen von PDF-Dateien vorerst auf die Kompressionsart JBIG2 verzichtet und ein anderer Kompressionsalgorithmus verwendet werden (JPEG, JPEG2000, CCITT). Die dadurch erhöhte Dateigrösse muss gegebenenfalls in Kauf genommen werden.

Die verschiedenen Quellen, insbesondere die Informatikstellen der Abliefernden Stellen, sollen entsprechend sensibilisiert werden, damit die Anzahl fehlerhafter Scan-Dateien in Zukunft reduziert werden können.

Da der Fehler irreversibel ist, gibt es auch keine Konvertierungsanleitung. Die einzige Möglichkeit den Fehler zu beheben, wäre das erneute einscannen des Originalpapieres.

[15] ISO 19005-2:2011, Document management — Electronic document file format for long-term preservation — Part 2: Use of ISO 32000-1 (PDF/A-2), S. v: "By itself, PDF/A does not necessarily ensure that the visual appearance of the content accurately reflects any original source material used to create the conforming file, e.g. the process used to create a conforming file might substitute fonts, reflow text, downsample images or use lossy compression. […] In addition, it is important for those organizations to implement policies and practices regarding the inspection of conforming files for correct visual appearance."

5 Liste der Testdokumente

5.1 A - Vorlagen

- A1_Scan_Testseite.pdf
- A2_Scan_Testseite_Arial.7.8.9.pdf
- A3_Scan_Testseite_Wiederholungen.pdf
- A4_Scan_Testseite_Arial7.pdf
- A5_Scan_Testseite_Arial5.pdf
- A6_TIFF_Normal.tif

5.2 B - Ergebnisse

- B11_Standard.pdf
- B12_Qualitativ_Hochwertig.pdf
- B13_Freigeben_und_Drucken.pdf
- B14_Archivieren.pdf

- B15_WC7346_JBIG2_arith_Normal.pdf

- B21_Schriftgroesse_Arial.7.8.9.pdf

- B31_je_1_Zeile.pdf
- B32_je_2_Zeilen.pdf
- B33_je_3_Zeilen.pdf
- B34_je_4_Zeilen.pdf

- B41_Fujitsu.pdf
- B42_Fujitsu_Arial.5.pdf

- B51_TIFF_PDF_lossless.pdf
- B52_TIFF_PDF_lossy.pdf
- B53_TIFF_PDF_CCITT.G4.pdf

- B61_Arial.7_Patch.pdf

- B71_WC7346_TIFF_Normal_1.tif
- B72_WC7346_TIFF_Normal_2.tif
- B73_WC7346_TIFF_Normal_3.tif

5.3 C - Markierungen / Auswertungen

- C11_Standard_683.775_Markierung.pdf
- C12_Qualitativ_Hochwertig_683.775_Markierung.pdf
- C13_Freigeben_und_Drucken_683.775_Markierung.pdf
- C14_Archivieren_Fehler_Markierung.pdf