Bericht zum Thema Dateiformat und Datenkomprimierung im Projekt "Konservierung Planarchiv" des StAZH

1 Projektvorgaben

Die Projektbeschreibung sieht folgendes vor:
"Die Sicherung der Pläne soll durch Mikroverfilmung auf Vollfichen in Farbe im Format 10,5 x 14,8 cm gewährleistet werden. Auf einer Vollfiche befinden sich eine Abbildung des Objekts, ein Massstab, ein Farbkeil, die vom StAZH definierten Metadaten in analoger Form und der Bitstream des entsprechenden Digitalisats inkl. Metadaten als 2D-Barcode.

Die Digitalisierung der Pläne soll in guter Qualität erfolgen, so dass auch Reproduktionen möglich sind (mind. 300 dpi). Die Daten sollen im Farbmodus RGB aufgenommen und als unkomprimierte TIF-Datei (bzw. in einem noch zu definierenden Format) gespeichert werden."

Die Projektvorgabe, sowohl ein hochaufgelöstes Bild (mind. 300 dpi), wie auch ein entsprechendes Digitalisat als 2D-Barcode auf einer Fiche unterzubringen, verlangt eine Datenkomprimierung beim Digitalisat. Eine Speicherung als TIFF-Datei, das nahe liegende Archivformat in diesem Falle, würde bei der im Projekt MONOLITH vorgesehen *bits-on-film* Speicherdichte etwa das Zehnfache der zur Verfügung stehenden Filmfläche benötigen. Damit wären etwa fünf Fichen pro Plan notwendig, was im Rahmen des Projektbudgets nicht möglich ist.

Durch die Projektvorgaben muss also für die Speicherung des Digitalisats auf dem Film ein komprimierendes Bildformat verwendet werden. Eine Komprimierungsrate von Faktor 10 kann nicht verlustfrei erreicht werden: Verlustfreie Komprimierung erreicht bei Bildern in der Regel höchstens einen Komprimierungsfaktor 2.

2 Verlustbehaftete Bildkomprimierung

Verlustbehaftete Bildkomprimierung widerspricht eigentlich dem Anspruch auf Aufbewahren eines Originals grundsätzlich, weil nach der Komprimierung das ursprüngliche Objekt nicht wieder hergestellt werden kann. Wenn wir den Fall aber genauer betrachten, sehen wir, dass die ursprüngliche TIFF-Datei im Falle eines Digitalisierungsprojektes sowieso nicht "das Original" ist, sondern eine durch Inkaufnahme von mindestens theoretischem Informationsverlust erstellte digitale Kopie. Die Auflösung beim Scannen (in diesem Falle 300 dpi) bestimmt diesen anfänglichen, irreversiblen und je nach Detaillierungsgrad sichtbaren Informationsverlust. Durch die Komprimierung entsteht in der Regel kein weiterer sichtbarer Informationsverlust, weil die Komprimierung nur die unterschiedliche Informationsdichte auf dem Ursprungsbild ausnutzt. Nicht reversibel ist das Wiederherstellen der ursprünglichen TIFF-Datei. Das sichtbare Bild, gewonnen aus der TIFF-Datei oder aus einer komprimierten Datei, ist aber in

diesem Falle identisch; in der Literatur wird das *visual lossless image compression* genannt.

So betrachtet wird verlustbehaftete Bildkomprimierung dort zum Problem, wo wir *digital born data* archivieren wollen, weil wir dort bei der Komprimierung einen Teil des Informationsgehaltes des Originals vernichten. Bei der Digitalisierung lässt sich dieser Informationsverlust aber grundsätzlich nicht umgehen, ist doch ohne Rasterung eine Überführung von analogen in digitale Bilder nicht möglich.

Grundsätzlich würden die besten Ergebnisse, das heisst am wenigsten Informationsverlust, mit einer möglichst grossen Auflösung beim Scannen und einer anschliessenden Komprimierung auf handhabbare Grösse erreicht. Damit können sichtbare Informationsverluste am besten verhindert werden. Der Aufwand beim Scannen, sehr grosse Temporärdateien und der Rechneraufwand für die Komprimierung setzen hier aber Grenzen. Diese Überlegungen sind im Rahmen dieser Empfehlung nicht weiterverfolgt und überprüft worden.

Wie beurteilen wir nun die Qualität oder den Qualitätsverlust? Die Daten des komprimierten Files werden mittels hochauflösenden bildgebenden Verfahren (z.B. hochauflösender [Inkjet o.ä.] Druck, Laser-Foto-Ausbelichtung) auf 100% bis 200% der Masse des Original-Dokumentes vergrössert und danach neben das Original gelegt. Der kritische Betrachter kann nun nach allfälligen Unterschieden suchen und sich ein Bild davon machen, ob diese Unterschiede detektierbar sind und – falls überhaupt vorhanden – ob sie akzeptabel wären oder nicht. Eine Betrachtung am Bildschirm ist grundsätzlich kritisch und sollte ein Herauszoomen von maximal „tatsächlichen Pixeln“ auf keinen Fall überschreiten (vgl. hierzu die praktischen Muster und Reproduktionen im Anhang).

3 Dateiformat

Bei der Speicherung des Digitalisats auf dem Film kommt der Wahl des Dateiformats eine grosse Bedeutung zu, soll das Digitalisat doch während der ganzen Lebensdauer des Films lesbar sein; eine Formatmigration ist hier grundsätzlich nicht möglich.

Alternativ zu TIFF kommen als komprimierungsfähige Bildformate JPEG, PNG und JPEG2000 in Frage. Eine archivische *best practice* Bewertung lässt uns auf JPEG2000 fokussieren, wir wollen aber dennoch die andern beiden Formate kurz betrachten:

JPEG wird im Allgemeinen, trotz hoher Verbreitung und ISO-Standardisierung, nicht als Archivformat betrachtet. Das Format hat kein grosses Entwicklungspotential mehr und wird längerfristig wahrscheinlich durch das technologisch bessere JPEG2000 abgelöst. Bei einem Komprimierungsfaktor von 10 sind bei JPEG bereits gewisse Komprimierungsartefakte sichtbar. Da dies aber sehr

stark von der Art des komprimierten Bildes abhängt, kann die Eignung von JPEG nur durch Tests mit Originaldaten abgeklärt werden.

PNG erreicht in der Regel verlustfreie Kompressionsraten von 1 zu 10, aber unter Reduktion der Farbtiefe. Es fragt sich, ob das mit der speziellen Natur der zu archivierenden Pläne vereinbar ist. PNG kann grundsätzlich als archivtaugliches Format gelten, auch wenn es in den Archiven nicht besonders häufig vertreten ist.

JPEG2000 ist das Format mit dem besten Komprimierungsvermögen, was Bilddaten betrifft. Eine Kompressionsrate von 1 zu 10 sollte problemlos möglich sein, ohne dass sichtbare Einbussen in Kauf genommen werden müssen.

JPEG2000 ist ein Format mit grossem Zukunftspotenzial. Die Formatspezifikation ist offen gelegt und es existieren mehrere *open source*-Implementierungen. Einziger Nachteil ist die geringe Verbreitung von Dateien in diesem Format. Der Grund dafür ist einerseits, dass die Unterhaltungsindustrie bis heute noch keine Fotoapparate entwickelt hat, die dieses Format unterstützen, und anderseits, dass auch Standardbetriebsysteme und Browser im Augenblick dieses Format nicht direkt anzeigen können. Im Hinblick auf eine *bits-on-film*-Lösung ist das kein Argument gegen JPEG2000, sollen die Daten doch in nächster Zukunft nicht eingelesen oder gar online genutzt werden.

Gewisse Features, welche die Stärke von JPEG2000 ausmachen, kommen aber im Falle von *bits-on-film* nicht direkt zum Tragen, so etwa das Auslesen unterschiedliche Auflösungen oder *thumbnails* aus einer Quelldatei. Andere Features, wie unterschiedliche Komprimierungsraten je nach Bildregion können aber bei der Natur der Plansammlung sehr wohl von Vorteil sein.

4 Testdaten

Für eine erste Untersuchung mit Daten aus dem Projekt stehen uns fünf gescannte Karten zur Verfügung (insgesamt 1.5 GB). Die besonders detailreiche Karte 003 ist sowohl in 300 wie auch in 400 dpi gescannt.

Dazu haben wir je drei komprimierte Versionen in JPEG2000 und in JPEG. Die gelb markierte Spalte (10proz) bezeichnet die aus ökonomischen Überlegungen notwendige Komprimierung auf durchschnittlich 10% der Ursprungsdateigrösse.

Dateiname	orig-tiffs	j2k-2proz	%	j2k-10proz	%	j2k-lossless	%
Karte_001.tif	142'265'452	2'937'100	*2.06%*	14'359'324	*10.09%*	70'561'833	*49.60%*
Karte_002.tif	139'427'672	2'983'411	*2.14%*	14'655'019	*10.51%*	72'099'940	*51.71%*
Karte_003.tif	334'627'832	8'158'634	*2.44%*	40'470'207	*12.09%*	200'126'629	*59.81%*
Karte_003_400dpi.tif	587'078'164	14'282'228	*2.43%*	71'049'906	*12.10%*	71'049'906	*12.10%*
Karte_004.tif	115'384'916	2'240'907	*1.94%*	11'048'961	*9.58%*	54'694'584	*47.40%*
Karte_005.tif	116'795'176	1'555'539	*1.33%*	7'558'413	*6.47%*	37'222'786	*31.87%*
Karte_006.tif	114'290'180	2'401'363	*2.10%*	11'757'167	*10.29%*	57'823'532	*50.59%*
Total in Byte	**1'549'869'392**	**34'559'182**	***2.23%***	**170'898'997**	***11.03%***	**563'579'210**	***36.36%***
Prozent Original	**100.00%**	**2.23%**	***2.23%***	**11.03%**	***11.03%***	**36.36%**	***36.36%***

Dateiname	orig-tiffs	jpg-2proz	%	jpg-10proz	%	jpg-best	%
Karte_001.tif	142'265'452	5'014'940	*3.53%*	9'773'757	*6.87%*	43'865'106	*30.83%*
Karte_002.tif	139'427'672	6'783'946	*4.87%*	12'837'587	*9.21%*	47'742'824	*34.24%*
Karte_003.tif	334'627'832	27'375'075	*8.18%*	47'629'080	*14.23%*	145'799'800	*43.57%*
Karte_003_400dpi.tif	587'078'164	43'177'214	*7.35%*	72'601'424	*12.37%*	256'830'109	*43.75%*
Karte_004.tif	115'384'916	4'035'949	*3.50%*	7'850'347	*6.80%*	34'288'574	*29.72%*
Karte_005.tif	116'795'176	1'925'091	*1.65%*	4'264'094	*3.65%*	18'781'599	*16.08%*
Karte_006.tif	114'290'180	5'552'060	*4.86%*	10'313'010	*9.02%*	38'411'294	*33.61%*
Total in Byte	**1'549'869'392**	**93'864'275**	***6.06%***	**165'269'299**	***10.66%***	**585'719'306**	***37.79%***
Prozent Original	**100.00%**	**6.06%**	***6.06%***	**10.66%**	***10.66%***	**37.79%**	***37.79%***

Die Testdaten als TIFF, JPEG2000 und JPEG mit unterschiedlichen Komprimierungsraten

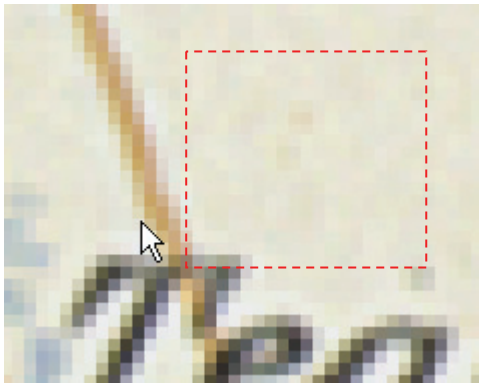
Im Anhang Ausschnitte aus den betrachteten Testdaten in unterschiedlicher Vergrösserung.

5 Beurteilung anhand der Testdaten

In der Originalansicht (100%) und bis zu einer vierfachen Vergrösserung (400%) lässt sich kein Unterschied zwischen den TIFF-Dateien und der Komprimierung auf 10% Dateigrösse in JPEG bzw. JPEG2000 feststellen. Wir können also hier von einer *visual lossless* Komprimierung sprechen (vgl. Muster aus dem Vorprojekt StAZH „Konservierung Planarchiv).

Bei der vierfachen Vergrösserung sehen wir aber bereits deutlich, dass die Auflösung schon beim Scannen zu einem Informationsverlust geführt hat und wir also im Falle der TIFF-Datei als Ergebnis des Scan-Prozesses nicht mehr von einer Originaldatei sprechen können. Das Original ist in diesem Falle die Karte im Archiv. Auf diesen Informationsverlust durch das Scannen hat die verlustbehaftete Komprimierung keinen Einfluss.

Erst bei einer sechsfachen Vergrösserung erkennen wir den Einfluss der Komprimierung, wenn wir einzelne Bildpunktfelder betrachten. Die unbedruckte Papierfläche in beiger Farbe wird z.B. geglättet (eigentlich wird hier ein Artefakt des Scannvorgangs korrigiert).



TIFF unkomprimiert

JPEG2000 komprimiert auf 10% Dateigrösse

Bei den betrachteten Testdaten lässt sich kein Unterschied bezüglich *visual lossless* zwischen JPEG und JPEG2000 erkennen. Mit JPEG2000 wird bei gleichen Einstellungen eine gleichmässigere Komprimierung erreicht, was die Dateigrösse betrifft; das kann bei der Gestaltung des Informationslayouts auf der Mikrofiche von Vorteil sein.

Bei noch höherer Komprimierung auf etwa 3% bis 6% der Ursprungsdateigrösse sind sowohl in JPEG wie auch in JPEG2000 Artefakte bereits in der Originalgrösse sichtbar. Die optimale/verlustfreie Komprimierung erreicht nicht die notwendige Dateigrössenreduktion.

6 Abschliessende Beurteilung

Die einzige verlässliche Aussage, die wir über die Langlebigkeit eines Formates machen können, ist die, dass ein Format mit einer hohen Verbreitung (d.h. eine Vielzahl von Dateien in diesem Format, die im Besitz oder Gebrauch von vielen Benutzern sind) eine relativ lange Lebenszeit besitzt. Vorausgesetzt der Markt ist nicht monopolisiert, werden die Dateibesitzer über längere Zeit die Nachfrage nach *rendering tools* für dieses Format sicherstellen.

Ein Beispiel: Die Generation der heute 30- bis 40-Jährigen, die ihre Bildersammlungen in JPEG angelegt und abgespeichert hat, wird nicht mehr auf ihre JPEG-Dateien verzichten wollen und diese auch nicht in ihrer Gesamtheit in ein neues Format konvertieren können. Damit ist diesem Format ein Überdauern von etwa 30 Jahren garantiert.

Unter diesem Gesichtpunkt müssten wir TIFF für die Speicherung wählen. Das geht aber aufgrund der fixen Projektparameter, nämlich dem verfügbaren Speicherplatz auf der Mikrofilmfiche nicht. Nächster Kandidat wäre JPEG, dann PNG, gefolgt von JPEG2000. PNG führt aber bei der gewünschten Komprimierung zu sichtbaren Komprimierungsartefakten, genauer Verlust von Farbtiefe, wir können hier nicht mehr von *visual lossless* Komprimierung sprechen. Es bleiben also JPEG und JPEG2000, die beide in diesem Falle die technischen Bedingungen erfüllen.

JPEG2000 besitzt den besseren Komprimierungsalgorithmus, was wir aber an den Testdaten nicht nachweisen können. Damit ist für JPEG2000 ein Zukunftspotential zu erwarten und wir können hoffen, dass es JPEG ablösen wird. JPEG2000 wird auch von andern Archiven als interessantes Format angesehen, sodass wir hoffen können, dass in den Archiven eine JPEG2000-Community entsteht, die das Verschwinden dieses Formates verhindern wird. Falls wider Erwarten JPEG2000 kein Massenerfolg beschieden ist, wird das Erhalten und Pflegen notwendiger *rendering tools* zur Aufgabe einer kleinen Benutzergruppe. Durch die Offenlegung der Spezifikation und durch *open source rendering tools* ist das aber technisch zu bewerkstelligen.

JPEG hingegen besitzt im Augenblick die viel grössere Verbreitung und wird aus den oben erwähnten Gründen die nächsten 30 Jahre als Format weiter, wenn vielleicht nicht mehr als Speicherformat attraktiv, so doch auf jeden Fall benutzbar sein.

Betrachten wir einen längern Zeitraum als 30 Jahre, was wir ja müssen im Falle von *bits-on-film,* hat aber wahrscheinlich JPEG2000 als kommendes Format die Nase vorne.

Wir schlagen also vor, unter Inkaufnahme eines kleinen Risikos JPEG2000 zu wählen, in der Hoffnung, dass sich dieses Format in naher Zukunft etablieren wird und wir dann von diesem Format eine längere Zukunftsperspektive erwarten können.

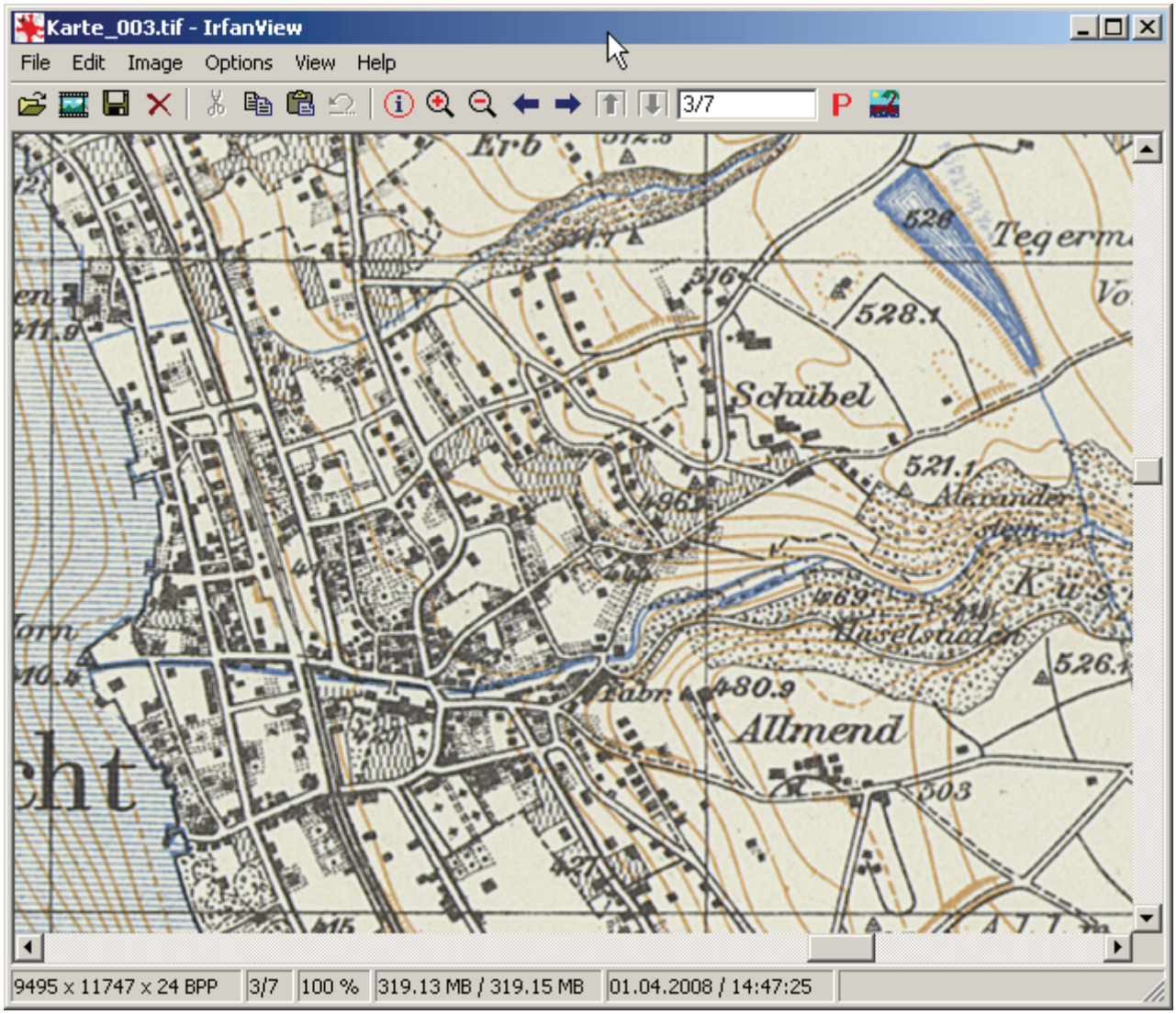
7 Zusammenfassung

Qualitätsbegriff: Die Originale sind handgezeichnet und gedruckt, gedacht für das Lesen mittels des menschlichen Auges im Massstab 1:1. Daraus sind die Ansprüche der Auftraggeberin, insbesondere bezüglich der Auflösung der Scans, abzuleiten. 300dpi entsprechen einer guten Qualität, die alle Information lesbar wiedergibt; von einer Faksimile-Sicherung kann jedoch nur bedingt gesprochen werden. Aus praktischen Gründen ist diese hundertprozentige Faksimile-Forderung aber auch nicht nötig. (Diese Beurteilung hat durch die Auftraggeberin selbst zu erfolgen).

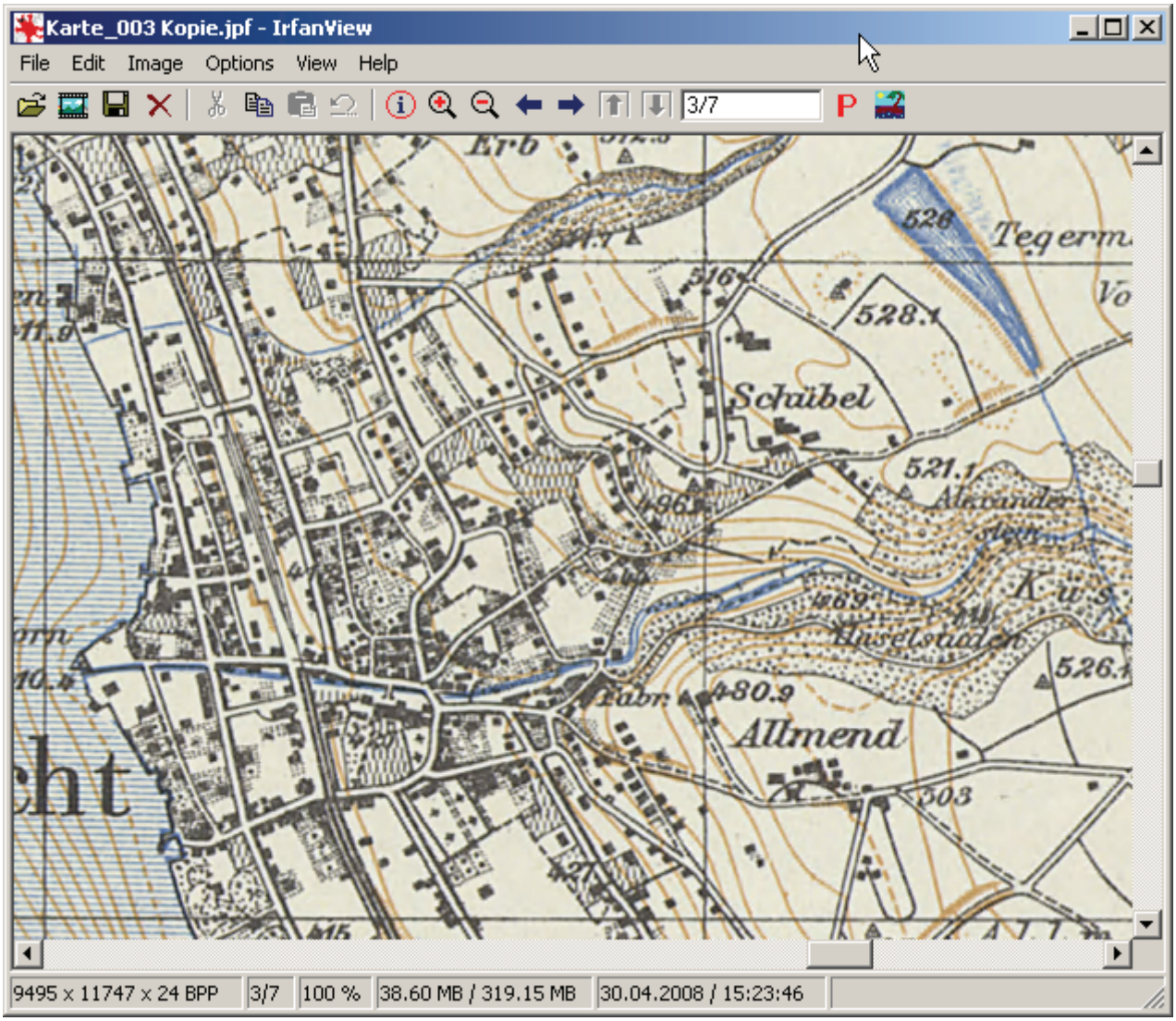
Dateiformat JPEG2000: Zur Wahl des Dateiformates sind vor allem Überlegungen zur Verbreitung eines Formates entscheidend. Wir schlagen also vor, unter Inkaufnahme eines kleinen Risikos, JPEG2000 zu wählen, in der Hoffnung, dass sich dieses Format in naher Zukunft etablieren wird und wir dann von diesem Format eine längere Zukunftsperspektive erwarten können.

Verlustbehaftete Komprimierung: Eine Kompressionsrate von 10% ist aufgrund der aktuell vorliegenden Muster vertretbar und kann mit JPEG wie auch mit JPEG2000 als *visual lossless* bezeichnet werden.

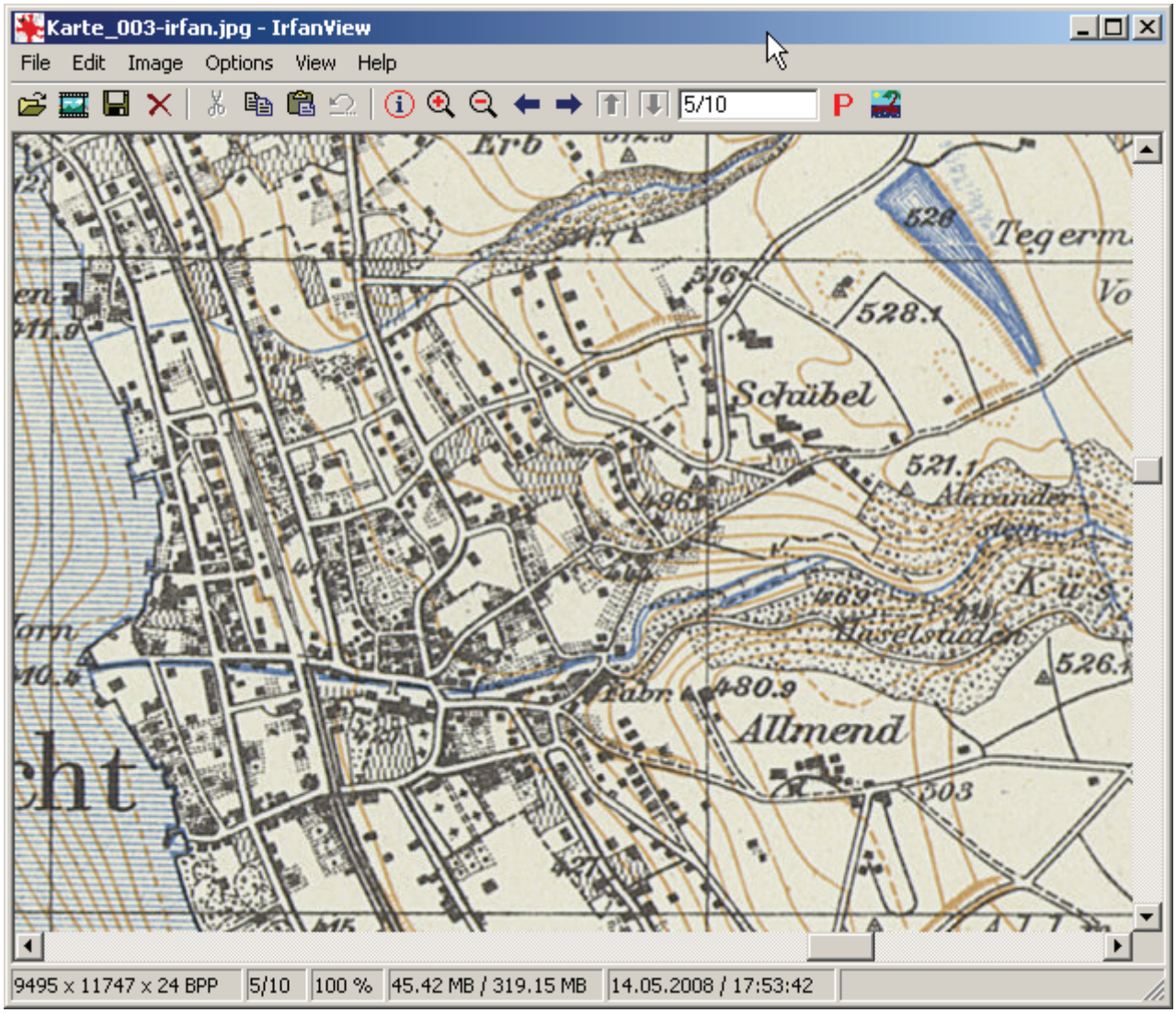
Testdaten



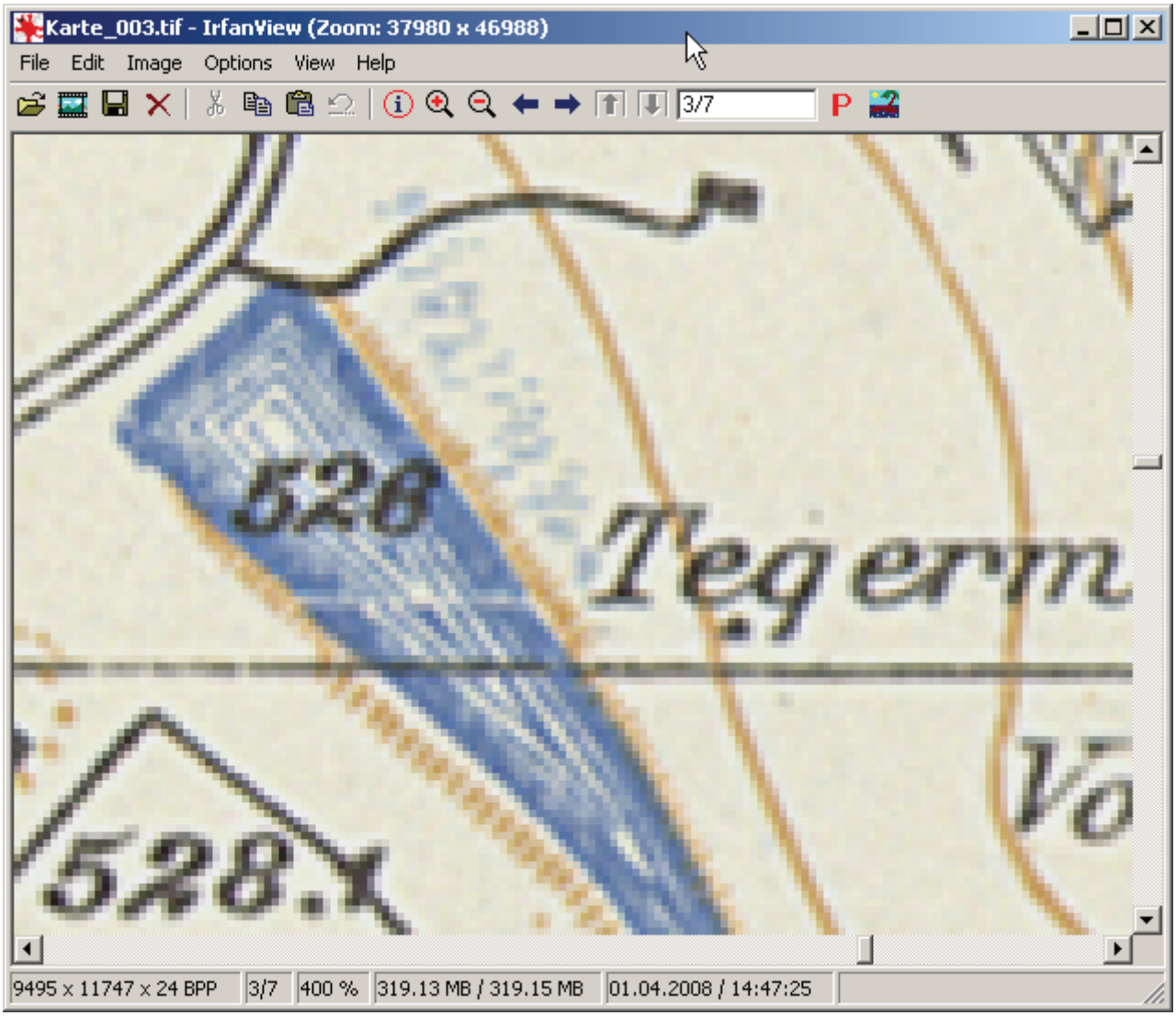
TIFF unkomprimiert, Ansicht 100%



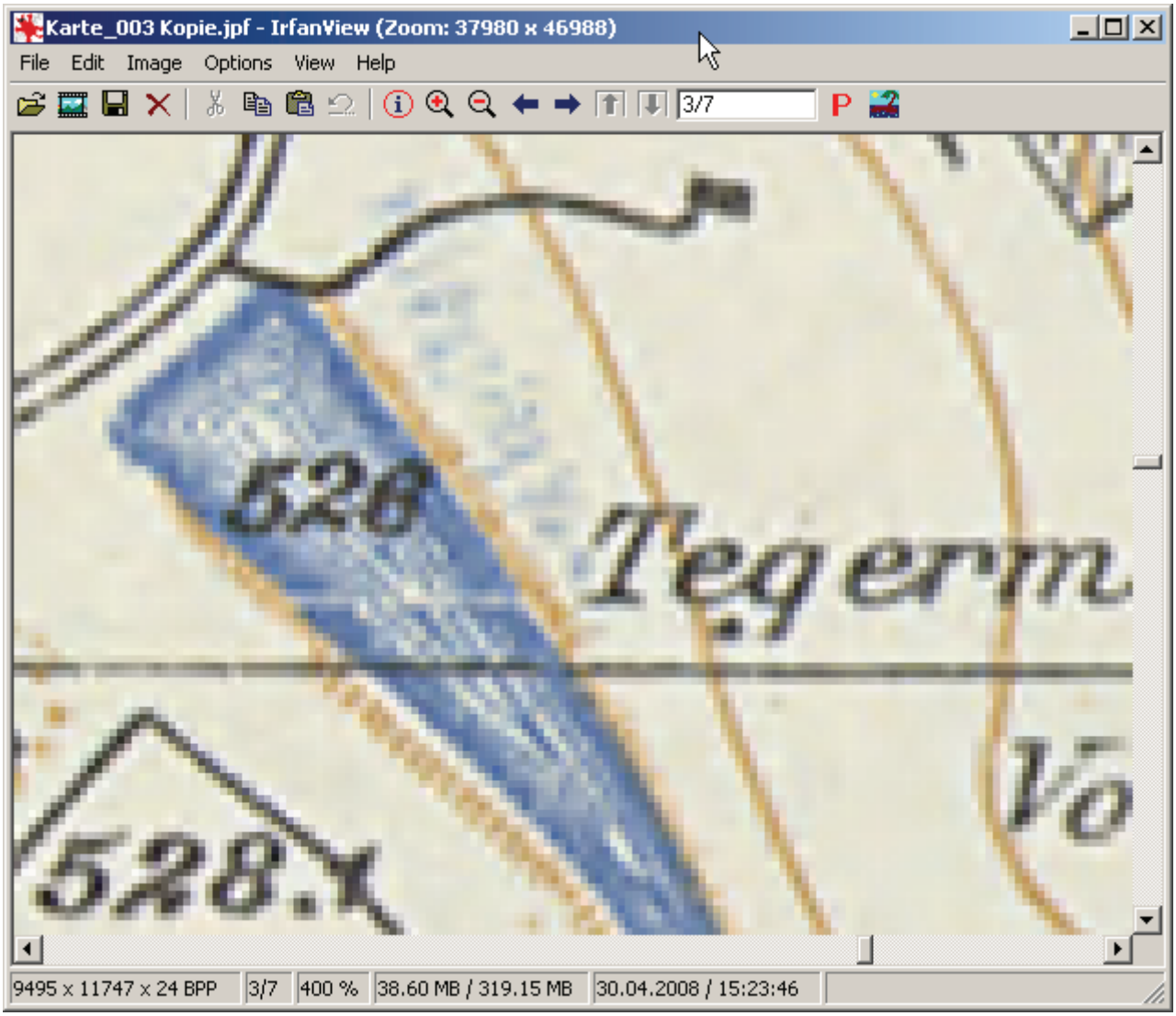
JPEG2000 komprimiert auf 10% Dateigrösse, Ansicht 100%



JPEG komprimiert auf 10% Dateigrösse, Ansicht 100%

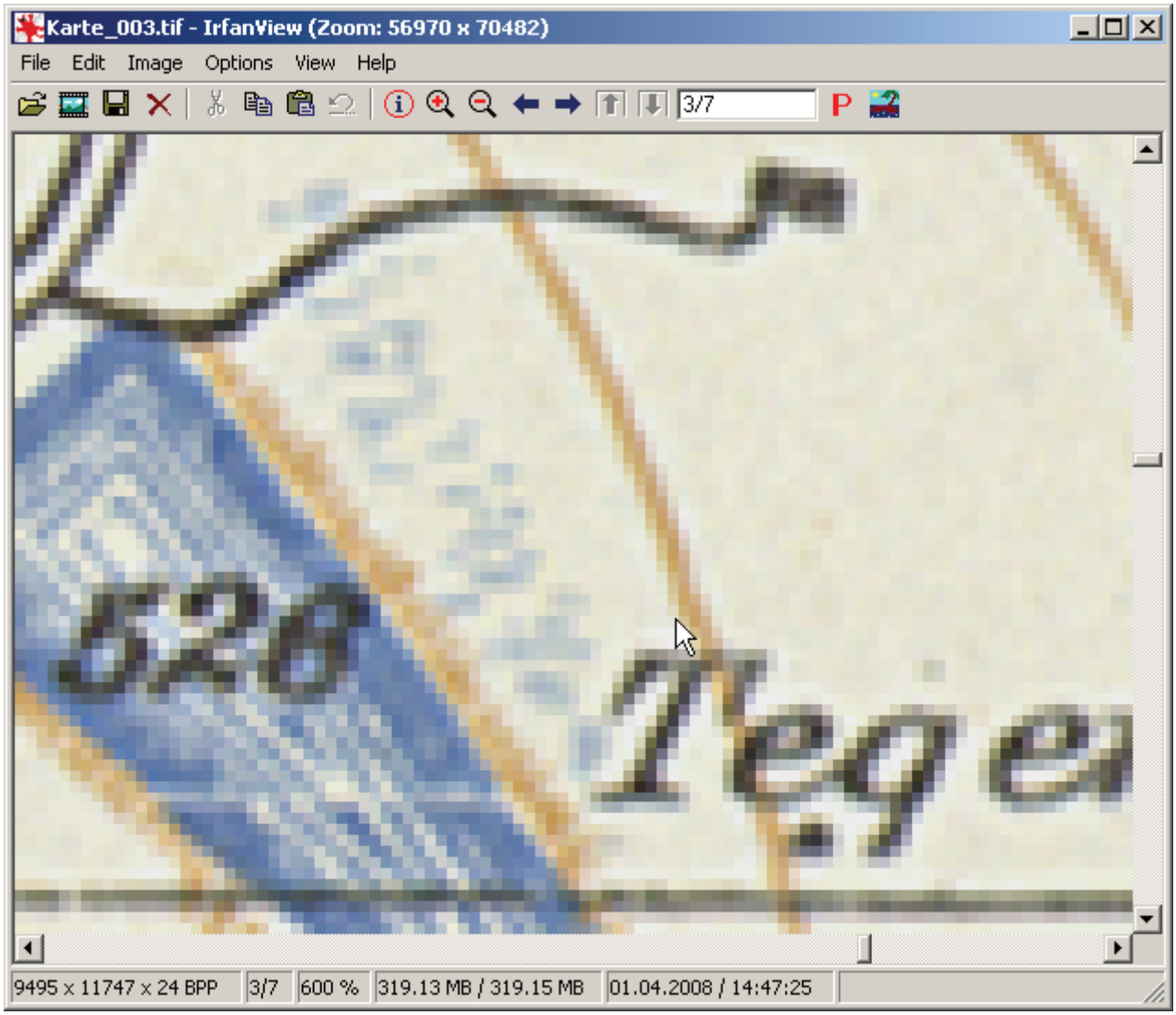


TIFF unkomprimiert, Ansicht 400%

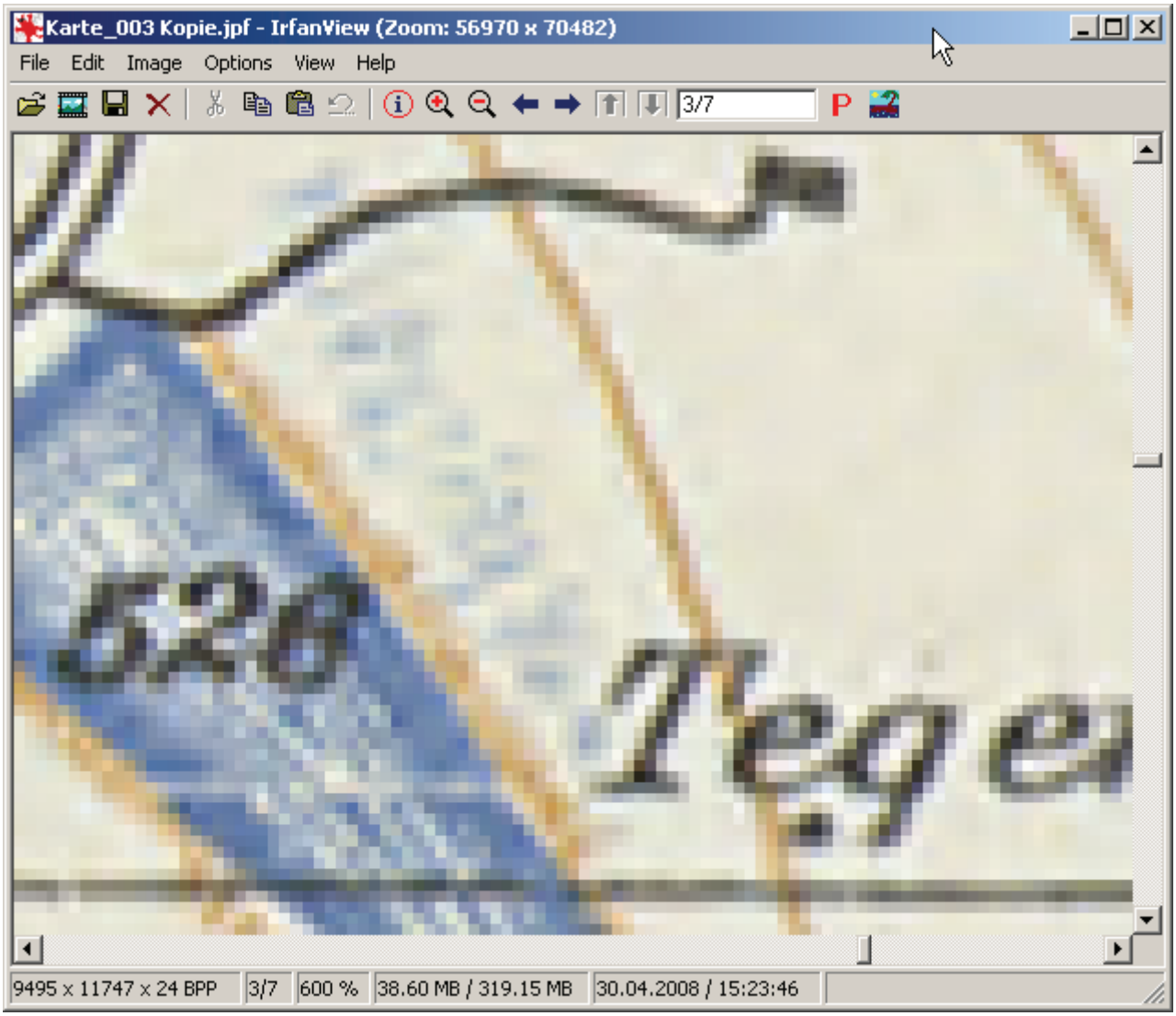


JPEG2000 komprimiert auf 10% Dateigrösse, Ansicht 400%

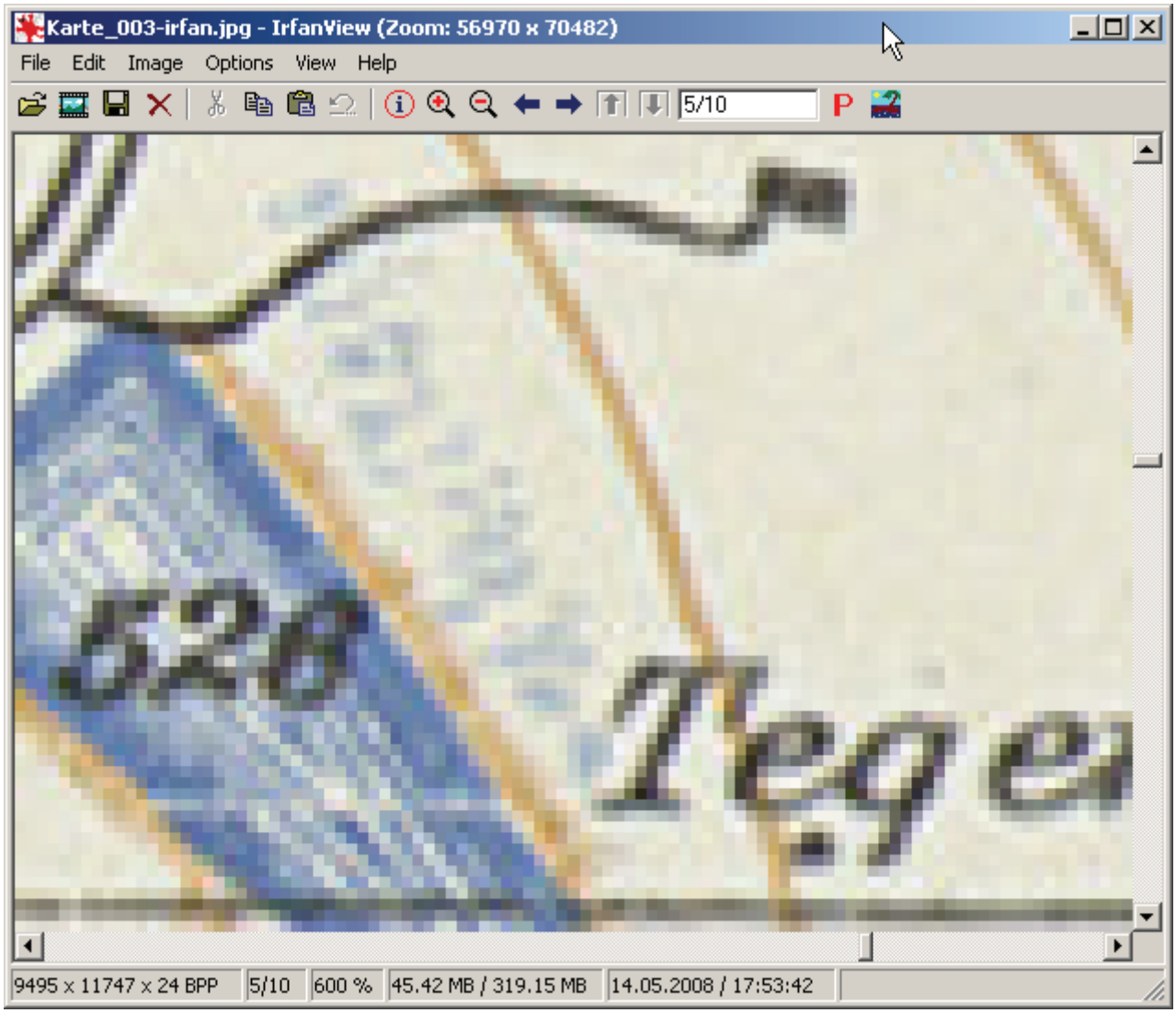
JPEG komprimiert auf 10% Dateigrösse, Ansicht 400%



TIFF unkomprimiert, Ansicht 600%



JPEG2000 komprimiert auf 10% Dateigrösse, Ansicht 600%



JPEG komprimiert auf 10% Dateigrösse, Ansicht 600%