

TIF Format (Tagged Image File Format) *Eine Preservation Planning Studie*

1	Einleitung	1
2	Analyse	1
2.1	Haupteigenschaften	1
2.2	Einschränkungen	2
2.3	Grundstruktur	2
2.4	Tags	3
2.5	Probleme in der Praxis	5
3	Beurteilung und KaD Preservation Planning Empfehlung	6

1 Einleitung

Das *Tagged Image File Format* (TIFF oder auch kurz TIF) ist ein Dateiformat zur Speicherung von Bilddaten. Es wurde ursprünglich von Aldus (1994 von Adobe übernommen) und Microsoft für die Farbseparation bei gescannten Rastergrafiken entwickelt. TIFF-Dateien besitzen den MIME-Typ `image/tiff` sowie die PUID `fmt/353`.

Eine genaue Beschreibung des Formats für Entwickler wird von Adobe kostenlos als PDF-Datei zur Verfügung gestellt. Die neueste Version ist 6.0 vom 3. Juni 1992¹. Sie wird ergänzt durch TIFF Technical Notes.

Die vorliegende Einführung in das Format ist in erster Linie eine Zusammenfassung aus Wikipedia, welche mit anderen Quellen, insbesondere der TIFF 6.0 Spezifikation, ergänzt wurde.

2 Analyse

2.1 Haupteigenschaften

TIFF kennt verschiedene Farbräume (*Photometric Interpretation*) und Algorithmen zur Datenkompression (*Compression*). Die meisten von ihnen sind verlustfrei (z. B. LZW, Lauflängenkodierung).

Einzelne Bildpunkte können bei TIFF aus beliebig vielen Einzelwerten (*Samples*) bestehen. *Samples* können neben dem Standardfall „ein Byte gleich ein *Sample*“ auch Teile eines Bytes einnehmen (z. B. 1, 2 oder 4 Bits) oder aus mehreren Bytes bestehen. Neben Ganzzahlen können auch Gleitkommazahlen als Bilddaten abgelegt werden.

Bilddaten werden in Gruppen von Pixelzeilen, so genannten *Stripes* (Streifen), oder als rechteckige *Tiles* (Kacheln) abgelegt. Die Speicherung erfolgt dabei für jeden Streifen beziehungsweise für jede Kachel unabhängig von den anderen, so dass Bildteile, je nach Wahl der Größe der Streifen beziehungsweise Kacheln, relativ schnell geladen werden können.

In einer Datei können mehrere Bilder abgelegt werden (*Multipage-TIFF*). Das können, müssen aber nicht, verschiedene Versionen desselben Bildes sein, zum Beispiel ein Vorschaubild („*Thumbnail*“) und das Originalbild.

¹ <http://partners.adobe.com/public/developer/tiff/index.html>

Die Möglichkeit zur Speicherung von Transparenz-Information (Alphakanal) existiert ebenfalls.

Programme wie Photoshop bieten an, TIFF-Dateien mit separaten Ebenen zu erstellen. Ferner gibt es die Möglichkeit, TIFFs mit einer Bildpyramide abzuspeichern. Dadurch sind innerhalb einer Datei mehrere Auflösungen des Bildes enthalten. Damit können zum Beispiel Layout-Programme oder Bildbetrachter eine kleine Vorschau des Bildes schneller anzeigen, da sie das Bild nicht in der vollen Auflösung laden müssen.

2.2 Einschränkungen

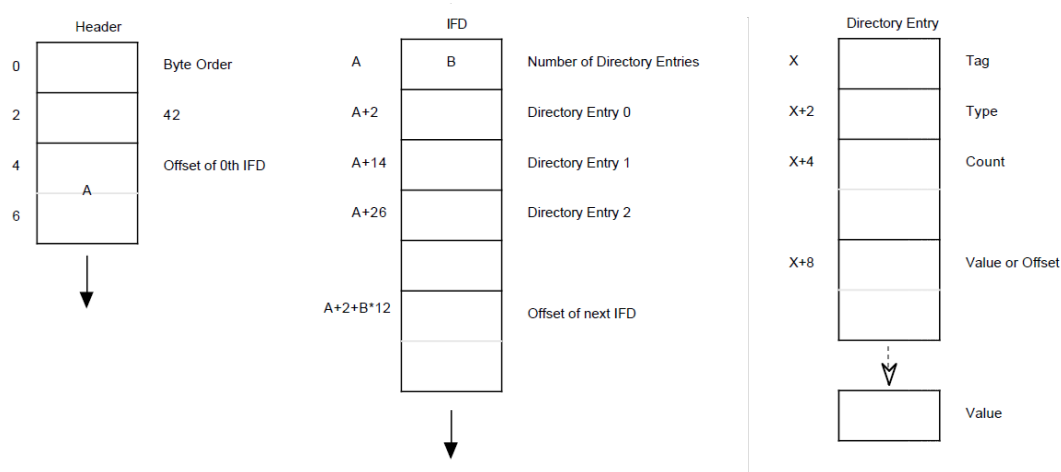
Grösster Nachteil von TIFF ist seine Komplexität. Die Vielfalt möglicher gültiger TIFF-Dateien kann nur schwer von einzelnen Programmen unterstützt werden. In der Spezifikation des Dateiformates ist deswegen unter dem Namen *Baseline TIFF* eine Untermenge gültiger TIFF-Dateien definiert (*Bilevel, Grayscale, Palette-color* sowie *RGB Full Color Images*), die jedes TIFF-fähige Programm verarbeiten können sollte.

Für Offset-Werte werden grundsätzlich 32 Bit verwendet. Das führt dazu, dass nur Stellen bis zu vier Gigabyte vom Dateianfang an referenziert werden können. Zur Zeit der Entwicklung von TIFF stellte das in der Praxis keine Einschränkung dar, in letzter Zeit fallen allerdings insbesondere in manchen wissenschaftlichen Disziplinen (z. B. Astronomie) sehr große Bilder an, die TIFF nicht speichern kann.

2.3 Grundstruktur

Die ersten acht Bytes einer TIFF-Datei bilden den *Header* und enthalten eine Signatur, die die *Byte Order* und die Magische Zahl 42 enthält, außerdem ein Offset zum ersten *Image File Directory* (IFD). Die gängigen Formaterkennungstools erkennen TIFF-Dateien anhand der ersten acht Bytes. Es ist jedoch nicht möglich eine Erkennung basierend auf den *Tags* zu erhalten.

Ein solches IFD gibt Informationen zu einem Bild in der TIFF-Datei und besteht aus einer Liste von *Tags*, einzelnen Informationseinheiten (*Directory Entry*). Ein solches *Tag* kann etwa die Breite des Bildes in Pixeln beschreiben oder den Namen der Software enthalten, die die TIFF-Datei erzeugt hat.



Jedes *Tag* hat eine eigene Nummer, die es kennzeichnet (z.B. Bildbreite: 256, *ImageWidth*), und einen *Type* (z.B. 16-Bit-Ganzzahlen, 32-Bit-Gleitkommazahlen, Zeichenketten u.v.m.). Es folgen die Angaben über die Länge des Datenbereiches sowie den Datenwert beziehungsweise den Zeiger auf Datenbereiche, sollte dieser

Grösser als 4 Byte sein. Die Anordnung der IFDs sowie der verschiedenen Daten ist nicht vorgeschrieben sondern durch die *Offsets* (Zeiger) gegeben.

Manche *Tags* müssen vorhanden sein (z. B. die Bildbreite), andere sind optional (z. B. der Name der Software). Das *Tag StripOffsets* (273) respektive *TileOffsets* (324) enthält für jeden *Stripes* oder *Tiles* den *Offset* zu diesem *Stripes* oder *Tiles*.

Eine Reihe von Anwendungen verwendet proprietäre *Tags*. Der Aufbau der Daten, die in einem solchen *Tag* gespeichert oder referenziert werden, ist dann meist nicht dokumentiert. Eine Nummer für eigene proprietäre *Tags* kann man bei Adobe beantragen.

Am Ende des IFD steht ein Offset-Wert, der das nächste IFD in der Datei referenziert, oder aber 0, falls das aktuelle IFD das letzte war. Auf diese Art können beliebig viele Bilder in der Datei untergebracht werden, solange die Gesamtmenge der Daten vier Gigabytes nicht übersteigt. Die *Baseline TIFFs* sind keine *Multipage-TIFFs* und enthalten entsprechend nur eine IFD.

2.4 Tags

Die TIFF-Spezifikation Version 6.0 kennt 74 *Tags*. Unzählige weitere *Tags* sind proprietär definiert; nicht alle davon wurden offiziell angemeldet. Die KOST hat 325 *Tags* aus der Spezifikation sowie aus vier weiteren Quellen zusammengetragen und in einer separaten Excel-Liste dokumentiert (TIFF_Tags_v1.1.xlsx). Es existieren aber weitere proprietäre *Tags*, deren genaue Anzahl unbekannt ist.

Die *Baseline TIFFs* definieren maximal 13 *Tags* als obligatorisch. Diese sind nachfolgend aufgeführt:

TIFF Tag					
Tag value		Tag Name	Type	Count	Number of values
hex	dez				
100	256	ImageWidth	SHORT / LONG	1	
101	257	ImageLength	SHORT / LONG	1	
102	258	BitsPerSample	SHORT	N	SamplesPerPixel
103	259	Compression	SHORT	1	
106	262	PhotometricInterpretation	SHORT	1	
111	273	StripOffsets	SHORT / LONG	N	StripsPerImage
115	277	SamplesPerPixel	SHORT	1	
116	278	RowsPerStrip	SHORT / LONG	1	
117	279	StripByteCounts	SHORT / LONG	N	StripsPerImage
11A	282	XResolution	RATIONAL	1	
11B	283	YResolution	RATIONAL	1	
128	296	ResolutionUnit	SHORT	1	
140	320	ColorMap	SHORT	3*N	3 * (2**BitsPerSample)

Die *Tags* zu den Farbräumen (262, *PhotometricInterpretation*) und Algorithmen zur Datenkompression (259, *Compression*) kennen selber noch *Subtags*.

Zu *Compression* existieren mindestens 29 verschiedene Verfahren, wovon lediglich 9 in der TIFF-Spezifikation Version 6.0 und deren Technical Notes beschrieben sind. Von deren 9 sind 3 als *Baseline* definiert. Nachfolgend werden nur die 9 Verfahren aus der TIFF-Spezifikation und deren Technical Notes aufgeführt:

TIFF Compression Tag (0103h / 259)						
Tag value		Compression scheme	Lossy / lossless	Description	Type of images	Usage & support
[hex]	[dez]					
1	1	Uncompressed	Lossless	Baseline TIFF	all	common
2	2	CCITT Group 3 1-Dimensional Modified Huffman run length encoding	Lossless	Baseline TIFF	black and white	common
3	3	<i>CCITT T.4 bi-level encoding</i>	Lossless	<i>TIFF ext.</i>	<i>black and white</i>	<i>common</i>
4	4	<i>CCITT T.6 bi-level encoding</i>	Lossless	<i>TIFF ext.</i>	<i>black and white</i>	<i>common</i>
5	5	<i>LZW (Lempel-Ziv & Welch algorithm)</i>	Lossless	<i>TIFF ext.; first defined in TIFF 5 as Baseline (1988)</i>	<i>all</i>	<i>common</i>
6	6	<i>JPEG (Obsolete 'old-style' JPEG, later overridden in Technote2)</i>	Lossy	<i>TIFF ext.; first defined in TIFF 6 (1992); Obsolete, should never be written.</i>	<i>continuous-tone</i>	<i>rare</i>
7	7	<i>JPEG ('new-style' JPEG)</i>	Lossy	<i>Technote2 overrides old-style JPEG compression; it is a TIFF ext.</i>	<i>continuous-tone</i>	<i>medium</i>
8	8	<i>Deflate Official version ('Adobe-style')</i>	Lossless	<i>Adobe Photoshop TIFF Tech. Notes; it is a TIFF ext.</i>	<i>all</i>	<i>medium</i>
8005	32773	PackBits compression, a.k.a. Macintosh RLE	Lossless	Baseline TIFF	all	common

Bei den *PhotometricInterpretation* existieren mindestens 14 verschiedenen Varianten, 8 sind in der TIFF-Spezifikation Version 6.0 beschrieben und 4 davon als *Baseline* definiert. Nachfolgend werden nur die 8 Varianten aus der TIFF-Spezifikation aufgeführt:

TIFF PhotometricInterpretation Tag (0106h / 262) Teil 1/2					
Tag value		Compression scheme	Description	Type of images	Usage & support
[hex]	[dez]				
0	0	WhitelsZero. For bilevel and grayscale images: 0 is imaged as white.	Baseline TIFF	bilevel and grayscale images	common
1	1	BlackIsZero. For bilevel and grayscale images: 0 is imaged as black.	Baseline TIFF	bilevel and grayscale images	common
2	2	RGB. RGB value of (0,0,0) represents black, and (255,255,255) represents white, assuming 8-bit components. The components are stored in the indicated order: first Red, then Green, then Blue.	Baseline TIFF	all	common
3	3	Palette color. In this model, a color is described with a single component. The value of the component is used as an index into the red, green and blue curves in the ColorMap field to retrieve an RGB triplet that defines the color.	Baseline TIFF	all	common
4	4	<i>Transparency Mask. This means that the image is used to define an irregularly shaped region of another image in the same TIFF file.</i>	<i>TIFF Ext.</i>	<i>all</i>	<i>medium</i>
5	5	<i>CMYK.</i>	<i>TIFF Ext.</i>	<i>all</i>	<i>medium</i>
6	6	<i>YCbCr</i>	<i>TIFF Ext.</i>	<i>all</i>	<i>medium</i>
8	8	<i>CIE L*a*b*</i>	<i>TIFF Ext.</i>	<i>all</i>	<i>medium</i>

Die Tags können mit speziellen Tools gelesen werden.

2.5 Probleme in der Praxis

Bei der vertieften Analyse der TIFF-Bestände in den Archiven sowie bei den einzelnen Testdateien wurden zwei weitere Probleme aufgedeckt.

Einerseits gibt es momentan nur einige Viewer, welche 16 BitsPerSample darstellen können. Die meisten Viewer rechnen diese auf 8 BitsPerSample runter, welches in einzelnen Fällen ein fehlerhaftes Bild ergeben kann. Bei 32 BitsPerSample sind die Anzahl möglicher Viewer sehr gering.

Andererseits ist die Dateigrösse ein weiteres Problem für die Viewer. Theoretisch sind Grössen bis 4 Gigabyte erlaubt und möglich. Praktisch können Dateien, welche grösser sind als ein Gibabyte, nur noch mit einzelnen Programmen gelesen werden.

3 Beurteilung und KaD Preservation Planning Empfehlung

Das Generieren einer TIFF-Datei ist relativ einfach, da das Format individuell angepasst werden kann. Dadurch bereitet dieses Format jedoch insbesondere bei der Benutzung sowie im Bereich der *Preservation Planning* grössere Probleme. Es existiert kein *TIFF-Viewer*, der alle TIFFs lesen (geschweige denn konvertieren) kann: Es gibt einfach viel zu viele Möglichkeiten, wie ein TIFF zusammengesetzt werden kann. Entsprechend ist eine Validierung bis ins Detail auch unmöglich.

Aus Preservation Planning Sicht ist es unerlässlich, genauer zu beschreiben, was ein "archivtaugliches TIFF" ist. Zudem muss im KaD eine Empfehlung bezüglich der TIFF-Varianten abgegeben werden, welche danach auch im Bereich der *Preservation Planning* beobachtet wird. Dafür wird ein TIFF-Set mit möglichst allen empfohlenen Varianten benötigt.

Die Empfehlung kann sich nicht nur auf die *Baseline TIFFs* beschränken. Die KaD Preservation Planning Empfehlung gibt vor, dass zu den *Baselines* zusätzlich auch die Kompressionsverfahren LZW (5), CCITT T.4 (3) und CCITT T.6 (4) sowie *Multipage-TIFFs* empfohlen werden. LZW zählte bereits früher zu den *Baseline TIFFs* und wurde in der Version 6.0 wegen des darauf bestehenden Patentes gestrichen; dieses Patent ist jedoch inzwischen abgelaufen. *Multipage-TIFFs* existieren, soweit der KOST bekannt, bereits in beträchtlicher Zahl in den Archiven. Zudem wird nebst den *Baseline*-Werten 1, 4 und 8 auch noch 16 BitsPerSample (pro Kanal) empfohlen sowie die Dateigrösse auf maximal 1GB eingeschränkt.

Auflistung aller empfohlenen Eigenschaften	
Datenkompression:	
- uncompressed	(1)
- CCITT 1D	(2)
- CCITT Group 3	(3)
- CCITT Group 4	(4)
- LZW	(5)
- PackBits	(32773)
Farbraum:	
- white is zero	(0)
- black is zero	(1)
- RGB	(2)
- palette color	(3)
BitsPerSample (pro Kanal):	
- 01	
- 04	
- 08	
- 16	
<i>Singelpage und Multipage</i>	
Streifen (→ Kacheln sind nicht empfohlen)	
Dateigrösse bis max. 1GB	