KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen

Ein Gemeinschaftsunternehmen von Schweizer Archiven

TIFF Korpus-Analyse

21. Tagung des Arbeitskreises AUdS,
Basel,
28.02.2017

Einleitung

http://kost-ceco.ch/cms/index.php?tiff-data-analysis_de

- KOST-Empfehlung
- ISO TIFF Recommendation
- TIFF-Bestände bei den beteiligten Archiven
- Prozess der Datenbereitstellung
- Prozess der Analyse
- Auswertung + Ergebnisse

Zusammenfassung: TIFF ist zurzeit eine offene Spezifikation von Adobe, jedoch kein *ISO-Standard*. Das Ziel des Vorhabens ist es, die theoretisch getroffenen Überlegungen, welche zu einer erweiterten, auf *baseline TIFF* basierenden *ISO Recommendation* führen sollten, auf einer fundierten Analyse echter archivischer Daten abzustützen.

KOST Empfehlung zu TIFF

http://kost-ceco.ch/cms/index.php?preservation_tiff_de

TIFF-Präzisierung:			
Bezeichnung & Tag-Nr. (Subtag)	TIFF Baseline V6 (Part1)	Erweiterung	Einschränkung
Datenkompression 259			
uncompressed (1)	X		
CCITT 1D (2)	X		
CCITT Group 3 (3)		X	
CCITT Group 4 (4)		X	
LZW (5)		X	
PackBits (32773)	X		
Farbraum 262			
white is zero (0)	X		
black is zero (1)	X		
RGB (2)	X		
palette color (3)	X		
BitsPerSample 258			
01		X	
04	X		
08	X		
16		X	
Anzahl Seiten	Singelpage	Multipage	
Bildaufbau	Streifen		
Dateigrösse			≤ 1GB

Die Erweiterungen betreffen

- einerseits den Bereich Datenkompression, wo sich seit Erscheinen von *Baseline TIFF* 1992 weitere Kompressionsalgorithmen etabliert haben
- anderseits die Erweiterung der Bittiefe auf heute bereits verbreitete 16 Bits Per Sample
- und eine Einschränkung der Dateigrösse im Hinblick auf verfügbare Viewer

ISO TIFF Recommendation

http://ti-a.org/

Von der KOST-Empfehlung zu einer ISO Recommendation

In Zusammenarbeit mit dem *Digital Humanities Lab* der Universität Basel (vormals *Image and Media Lab IML*) und der Universität Girona verfolgen wir das Projekt einer Standardisierung von TIFF im Hinblick auf Archivierung und die Digitalisierung in Gedächtnisinstitutionen. In Analogie zu PDF soll ein TIFF/A auf der Basis von *Baseline TIFF v6* spezifiziert und anschliessend wenn möglich als *ISO Standard* hinterlegt werden.

Das Projekt besteht aus zwei Teilen, einerseits der Spezifizierung von TIFF/A auf der Basis von *Baseline TIFF v6* als Basis für den Umgang mit bestehenden Bildsammlungen in diesem Format und andererseits einer *Best-Practice*-Spezifikation als Anleitung für die Digitalisierung analoger Bestände, bzw. als Vorgabe für die Fotografie mit dem Ziel, optimal archivtaugliche Daten zu generieren.

Ein weiterer indirekter Nutzen des Projektes ist der, dass die bei Adobe etwas verstreut abgelegten Dokumente zu TIFF neu bei ISO integral und langfristig hinterlegt sein würden.

Analyse bestehender TIFF-Bestände

Grundlage: eine Analyse bestehender TIFF-Bestände

Basel-Stadt		
Typ	Anzahl	Grösse
2000	1'950	
2001	2'300	
2002	200	
2003	100	
2004	10'000	
2005-2015	750'000	
Total:	764'550	12.4 TB

Bundesarchiv		
Typ	Anzahl	Grösse
Archivtiffs:	1.7 Mio	5-6 TB
Digitalisate:	7300	460 GB
Digitalisate 3te:	14 Mio	6 TB
Total:	15 Mio	11-12 TB

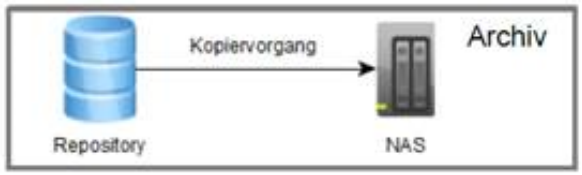
St. Gallen		
Typ	Anzahl	Grösse
Total:	870'000	12.83 TB

mittlere Grösse von 18,2 MB
grösste Datei 3,26 Gigabyte

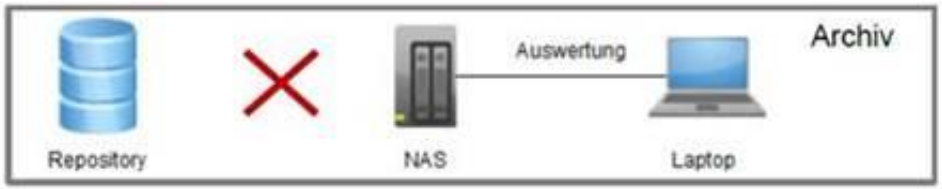
Dateien grösser 3GB:	2
Dateien >1 <3 GB:	2
Dateien 500M-1000M:	50
Dateien 100M-500M:	5000
Dateien <100M:	750000

Analyse Prozess

1. **Kopiervorgang**: Die Daten werden vom Archivrepository auf ein NAS oder eine USB-Platte kopiert



2. **Analyse**: Ausführen der eigentlichen Analyse ohne Verbindung mit dem Archivrepository



3. **Auswertung** in einem von der Archivinfrastruktur völlig unabhängigen, nachgeordneten Prozess

http://kost-ceco.ch/cms/index.php?tiff-data-analysis_de
https://github.com/KOST-CECO/TiffAnalyseProject

Analysemodule und -programme

- Eine simple Formaterkennung mit file erkennt falsch gelabelte Dateien.
- Die Validierung mit JHOVE ermittelt die grundlegende Struktur der TIFF-Datei. Wichtig sind hier Status und InfoMessage.
- DPF-Manager, Alternative zu JOHVE aus dem Preforma-Projekt.
- TIFF Tag-Extraktion mit tiffhist vom DHLAB. Das C++-Programm extrahiert alle Tags in eine CSV-Tabelle.
- checkit_tiff , ein conformance checker for baseline TIFFs, wurde von der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden entwickelt.
- Mit ExifTool werden eingebettete EXIF- und XMP-Metadaten extrahiert.
- Mit ImageMagick wird für jede Datei ein sehr kleines Thumbnail generiert. In diesem Schritt wird die Payload bzw. Bitmap der TIFF-Datei untersucht.

Einschränkungen bei der Analyse

Es hat sich schnell gezeigt, dass die verwendeten Tools ganz unterschiedliche Rechenzeiten pro Datei erfordern. Hier die Rechenzeiten pro Tool über 1000 Dateien unterschiedlicher Grösse (mittlere Grösse ~5.5 MB), verglichen mit *tiffhist*:

Tool	Sec	Prozent	Faktor
tiffhist	257	100%	1.0
dpf-manager	257	100%	1.0
file	266	104%	1.0
exiv2	267	104%	1.0
exif	335	130%	1.3
checkit_tiff	503	196%	2.0
jhove	697	271%	2.7
ImageMagick	3424	1332%	13.3
Total	6006	2337%	23.4

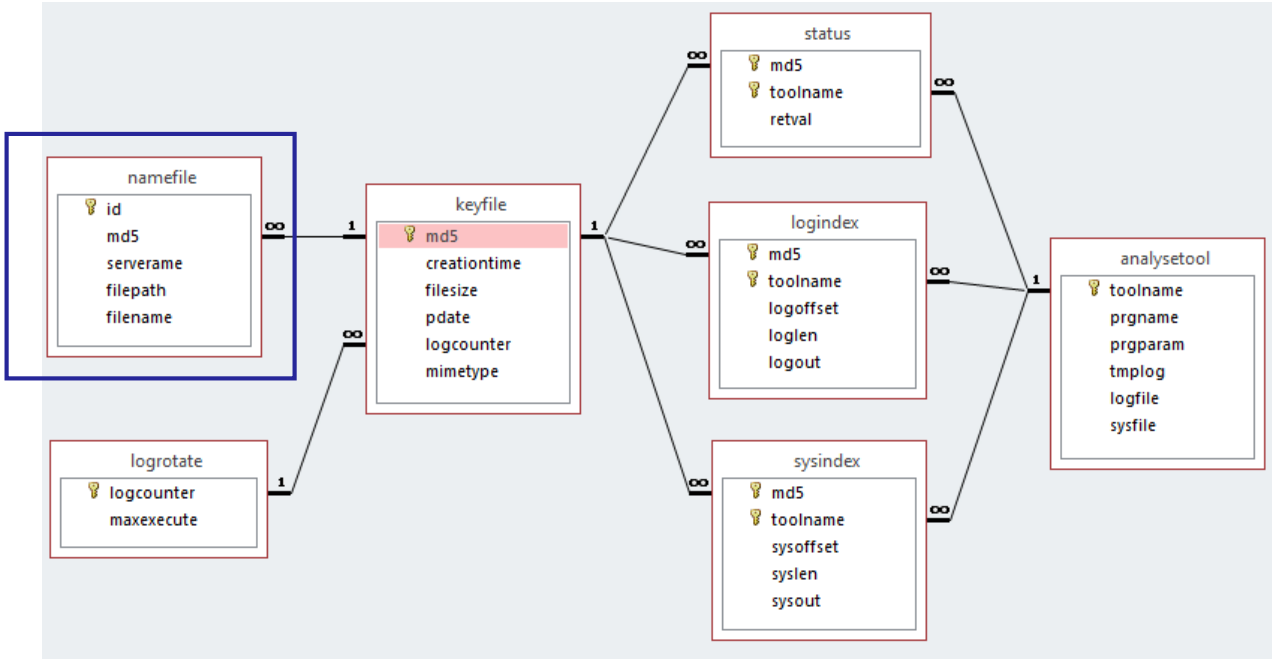
Dieser Umstand hat uns bewogen, *ImageMagick* nur über einem sehr kleinen Teilbestand und *JHOVE* nur etwa über der Hälfte der Dateien ausführen zu lassen.

Auswertung der Logdaten

- Analysemodule und -programme erzeugen insgesamt etwa 35 GB bzw. 150 Mio Zeilen Log-Informationen. Die gewonnen Informationen sind programmspezifisch äusserst unterschiedlich: Einzelwert, CSV-Liste, XML-Datei, zeilenorientierte Logdatei.
- Aus Effizienzgründen werden die Informationen nicht beim Ausführen der Programme analysiert oder aufbereitet, sondern nur gespeichert. Damit ist eine zeitversetzte Offline-Auswertung möglich.
- Damit eine Zuordnung von untersuchten Dateien und verschiedenen Analyseergebnissen möglich ist, wird jeder Analyseschritt in einer Datenbank festgehalten.
- Dateiname und Dateipfad werdenauf Wunsch der beteiligten Archive in der Datenbank und den Logdateien anonymisiert.

Daten und Programmierung

Die Logdaten und die Informationen zum Programmverlauf sind in einer Sqlite-Datenbank festgehalten. Die Tabelle *namefile* bleibt im Archiv, sodass auch später eine De-Anonymisierung möglich ist. Für das Überwachungs- und Steuerungsprogramm wurde Golang gewählt: http://kost-ceco.ch/cms/index.php?tiff-data-analysis_de



Beispiele für die Auswertung

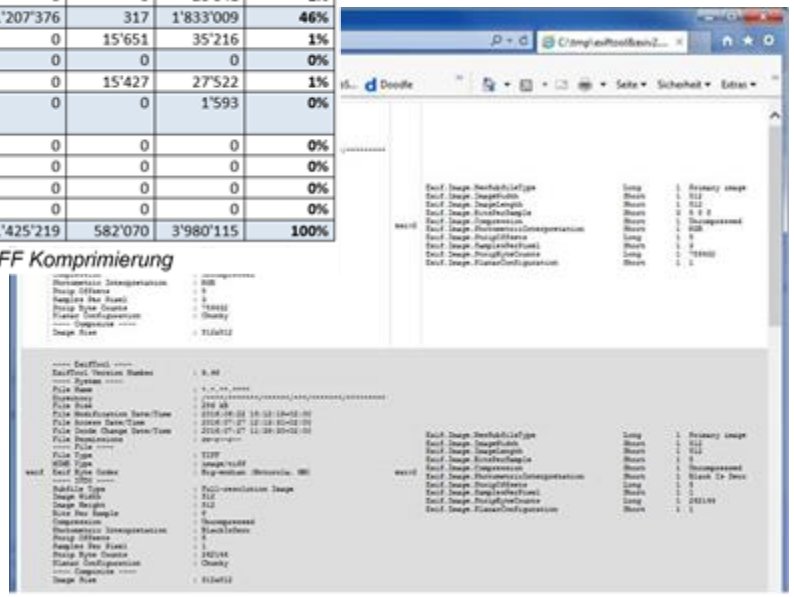
Diese Auswertung kann und soll nicht durch die KOST-Geschäftsstelle oder die Uni Basel alleine geleistet werden. Deshalb publizieren wir die Rohdaten und Beispiele für die Auswertung auf unserer Website: http://kost-ceco.ch/ftp_space/TIFF-Analyse

Compression	Basel-Stadt	Bundesarchiv	Bundesarchiv extern	St. Gallen	Total	Prozent
none	732'696	560'956	217'843	550'675	2'062'170	**52%**
CCIT 1D	0	564	0	0	564	**0%**
Fax Group 3	0	20'041	0	0	20'041	**1%**
Fax Group 4	22'745	602'571	1'207'376	317	1'833'009	**46%**
LZW	31	19'534	0	15'651	35'216	**1%**
old JPEG	0	0	0	0	0	**0%**
JPEG	0	12'095	0	15'427	27'522	**1%**
Adobe De-flate	0	1'593	0	0	1'593	**0%**
JBIG bw	0	0	0	0	0	**0%**
JBIG color	0	0	0	0	0	**0%**
Pack Bits	0	0	0	0	0	**0%**
other	0	0	0	0	0	**0%**
Summe	755'472	1'217'354	1'425'219	582'070	3'980'115	**100%**

Beispiel: Verteilung TIFF Komprimierung

```
sqlite> select * from sysindex;
md5|toolname|sysoffset|syslen|sysout
85cc801d15a99ac55a02bb937e212347|jhove|0|1700|Jhov
Date: 2016-07-28 14:38:51 MESZ
RepresentationInformation:
/**********/*********/*******/*************.***
  ReportingModule: TIFF-hul, Rel. 1.5 (2007-10-02)
  LastModified: 2005-04-15 08:22:38 MESZ
  Size: 6528508
  Format: TIFF
  Version: 5.0
  Status: Well-Formed and valid
  SignatureMatches:
   TIFF-hul
  MIMEtype: image/tiff
  Profile: Baseline grayscale (Class G), TIFF/IT-MP (ISO 12639:1998), DLF
Benchmark for Faithful Digital Reproductions of Monographs and Serials:
grayscale and white
  TIFFMetadata:
   ByteOrder: little-endian
   IFDs:
    Number: 1
    IFD:
     Offset: 6527508
     Type: TIFF
```

Beispiel: Auswertung JHOVE



Beispiel: Vergleich zweier Tools (exif und exiv2)

Hands-On Logfile

Index of /ftp_space/TIFF-Analyse/

Name	Last modified	Size	Description
Parent Directory	14-Feb-2017 15:59	-	
TIFF-Analyse Projektbeschreibung v1.0.pdf	20-Feb-2017 17:47	752k	
log.tgz	13-Feb-2017 14:01	1214236k	
tap.sql.gz	13-Feb-2017 14:03	3386268k	

Proudly Served by LiteSpeed Web Server at kost-ceco.ch Port 80

Section 8: Baseline Field Reference Guide

BitsPerSample

Number of bits per component.

Tag = 258 (102.H)

Type = SHORT

N = SamplesPerPixel

Note that this field allows a different number of bits per component for each component corresponding to a pixel. For example, RGB color data could use a different number of bits per component for each of the three color planes. Most RGB files will have the same number of BitsPerSample for each component. Even in this case, the writer must write all three values.

Default = 1. See also SamplesPerPixel.

```
find . -name "*tiffhist*.log" -exec cut -f 4 {} ; | grep "BitsPerSample" >out.txt
find . -name "*tiffhist*.log" -exec cut -f 5 {} ; | grep "BitsPerSample" >>out.txt
sort -u out.txt
258$BitsPerSample:1
258$BitsPerSample:2
   ...
grep ":8" out.txt | wc -l
 1'813'688
grep ":16" out.txt | wc -l
    93'301
grep "258$BitsPerSample" *tiffhist* | wc -l
 3'959'776
```

Hands-On Database

```
.mode html
SELECT
--   sys1.md5,
     sys1.toolname,
     sys1.sysout,
--   sys2.md5,
     sys2.toolname,
     sys2.sysout
FROM
     (SELECT md5, toolname, sysout FROM sysindex WHERE toolname = "exif") sys1
INNER JOIN
     (SELECT md5, toolname, sysout FROM sysindex WHERE toolname = "exiv2") sys2
ON
     sys1.md5 = sys2.md5;
```