Preservation-Planning   
*TIFF-Analyse – rollende Planung*

**Inhalt**

[1 Planung 1](#_Toc447115521)

[2 Vorgehen in einzelnen Schritten 1](#_Toc447115522)

[3 Angaben von den Archiven 2](#_Toc447115523)

[4 Entwicklung 3](#_Toc447115524)

[4.1 Infrastruktur und Programmiersprachen 3](#_Toc447115525)

[4.2 Analyse Programm 3](#_Toc447115526)

[4.2.1 Korpus Initialisierung 3](#_Toc447115527)

[4.2.2 Korpus Analyse 4](#_Toc447115528)

[4.2.3 Datenmodell 4](#_Toc447115529)

[4.2.4 Skript Programmierung 6](#_Toc447115530)

[5 Analysemodule 6](#_Toc447115531)

[6 Auswertung 7](#_Toc447115532)

[7 Testdaten 7](#_Toc447115533)

# Planung

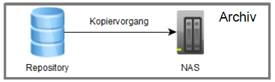
Wir haben folgendes Vorgehen mit dem *Digital Humanities Lab Basel* bezüglich Analyse von TIFF Dateien abgeklärt:

1. Jhove Analyse ermittelt grundsätzlich korrekter Aufbau der Datei, und Information über Alignement, Order und Typisierung der Tags und generiert das Jhove LOG
2. Ein Programm vom *Digital Humanities Lab (DHLAB)* liest sequenziell alle Tag, Tagnummern Tag Typen und Taginhalt und schreibt die Werte pro File in eine Datenbank (SQLite oder mySQL)
3. Mit EXIFTool werden allenfalls vorhanden EXIF und IPCT Metadaten ausgelesen und ebenfalls in die Datenbank geschrieben.
4. Als letztes wird optional ein *Thumbnail* generiert und in eine Datei geschrieben, damit kann die Integrität der Bitmap validiert und Bitmap Fehler einfach erkannt werden (z.B. alles Schwarz). Offen ist dabei, ob wir das Thumbnail LOG mitnehmen können oder ob die Auswertung nur im Archiv erfolgen kann.

Zur externen Auswertung wird das Jhove LOG und die TIFF-TAG/Metadaten Datenbank benötigt, das Thumbnail LOG muss allenfalls im Archiv ausgewertet werden. Das Thumbnail LOG wird nur genötigt, wenn im Jhove LOG oder in der TIFF-TAG/Metadaten Datenbank Anomalien festgestellt werden um.

# Vorgehen in einzelnen Schritten

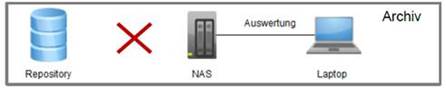
* Das Programm vom DHLAB muss vom Team von Herr Rosenthaler noch angepasst werden.
* Parallel dazu werden wir hier in der Geschäftsstelle eine Testinstanz mit unserer Hardware und eigenen Beispieldateien aufbauen und dort alle Komponenten testen können.
* In dieser Phase zeigen wir auch gerne den Beteiligten wie alles funktionieren soll. Wir hoffen, dass wir Ende Februar mit allen Vorbereitungen soweit sind.
* Sobald die Frage der Verfügbarkeit eines Raids in den beteiligten Archiven geklärt ist, können wir aber schon mit dem Kopieren der Dateien beginnen.   
  Dafür muss auf einem archivinternen Rechner mit Zugang zum Repository und dem NAS ein Kopierprogramm im Hintergrund laufen können (z.B. robocopy vom Microsoft)



# Angaben von den Archiven

Von Seiten der Archive benötigen wir nun noch folgende Angaben:

* Wer ist unser Ansprechpartner in diesem Projekt (BAR Marcel, StABS Markus, StASG ?)
* Grobes Mengengerüst: Anzahl Dateien / Speicherplatz
* Soll das NAS (10 TByte Raid 1) von KOST/DHLAB zur Verfügung gestellt werden oder kann das Archiv ein NAS für 4 Mnt. ausleihen?
* Soll der Rechner für die Auswertung von KOST/DHLAB zur Verfügung gestellt werden oder kann das Archiv ein Computer für 4 Mnt. ausleihen?
* Müssen die Speicherplatten im Raid nach der Auswertung im Archiv bleiben, wenn das NAS von KOST/DHLAB stammt oder reicht es die Daten mit einem sog. Shredder Programm zu überschreiben?



# Entwicklung

*Die einzelnen Aufgaben (Sitzung vom 15. März)*

1. Loop durch alle Verzeichnisse auf dem RAID Server mit Kopieren jeweils einer TIFF Datei vom RAID auf den Auswertungsrechen (mit Restartfähigkeit) KOST
2. Ein Hash Wert (z.B. md5) wird für die Datei berechnet um Doubletten zu erkennen. Der Hash Wert dient als Schlüssel in der DB  KOST
3. Die Jhove Analyse ermittelt grundsätzlich korrekter Aufbau der Datei und Information über Alignement, Order und Typisierung der Tags und schreibt das Jhove LOG. Der LOG-Offset wird zur schnelleren Analyse in die DB geschrieben.  KOST
4. Ein Programm vom Digital Humanities Lab (DHLAB) liest sequenziell alle Tag, Tagnummern Tag Typen und Taginhalt und schreibt die Werte in die DB
5. Mit EXIFTool werden allenfalls vorhanden EXIF und IPCT Metadaten ausgelesen und ebenfalls in die Datenbank geschrieben.  KOST
6. Als letztes wird optional ein Thumbnail generiert und in eine lokale Datei geschrieben, damit kann die Integrität der Bitmap validiert und Bitmap Fehler einfach erkannt werden (z.B. alles Schwarz). Der Thumbnail Name wird ebenfalls in der DB festgehalten. Offen ist dabei, ob wir das Thumbnail LOG mitnehmen können oder ob die Auswertung nur im Archiv erfolgen kann.

## Infrastruktur und Programmiersprachen

Vorschlag der Geschäftsstelle:

*Datenbank*: SQlite Vorteil, kein Server, keine Administration<https://www.sqlite.org/>

*Loop-Programm*: Golang, Vorteil kompilierte Sprache, auf allen Plattformen verfügbar, API zu SQLite und einfacher als C/C++ <https://golang.org/>

*Arbeitsbereich*: GitHub <https://github.com/KOST-CECO/TiffAnalyseProject>

Vorschlag Makrus:

Linux als Betriebssystem auf den Auswertungsrechnern: *Linux Mint 17.3*

## Analyse Programm

Das Analyse Loop Programm liest alle TIFF Dateien (Korpus) vom NAS und führt mit der jeweils gelesenen Datei mehrere Analyseschritte durch aufrufen externen Programme aus. Der Loop Prozess ist zweiteilig und besteht aus sein Initialisierung der Prozessdatenbank und der eigentlichen Analyse.

### Korpus Initialisierung

Ein erster Initialisierungsschritt erstellt die Datenbank und schreibt für jede TIFF Datei einen Eintrag mit dem Pfad und Dateinamen als Schlüssel. Argumente sind "Datenbank" und "Datei Root". Die Initialisierung kann mehrfach aufgerufen werden und fügt so neue Verzeichnispfade zur Datenbank hinzu.

Damit für eine spätere Auswertung ausserhalb der Archive problemlos gearbeitet werden kann, werden Dateinamen und Dateipfad, welche allenfalls Rückschlüsse auf den Inhalt der Dateien erlauben, in einer separaten Tabelle gehalten. Als gemeinsamer Schlüssel für alle spätere Auswertung wird der MD5 Schlüssel der TIFF Datei verwendet. Weil zum Berechnen des MD5 Schlüssels die ganze Datei gelesen werden muss, wird diese Berechnung erst im Analyseschritt vorgenommen.

### Korpus Analyse

Im Analyse Fall werden die Dateieinträge in der Datenbank abgearbeitet und als erstes ein MD5 Schlüssel berechnet. Damit können Doubletten im Dateisystem erkannt werden. Anschliessend werden die Analyse Tools im Kommandozeilen Modus aufgerufen.

Durch die Verwendung einer Datenbank ist jederzeit ein Abbrechen und wieder Starten der Analyse möglich.

* Dem Analyse Tools im Kommandozeilen Modus wird der Dateipfad und der Pfad zur Logdatei übergeben.
* Kann das Analyse Tools kein Logfile im Append Modus öffnen, kann der Log Output vom Loop Programm entweder an die Logdatei angehängt oder in die Datenbank geschrieben werden.
* Der aktuelle Offset der Log Datei wird in der Datenbank gespeichert.
* Der Exit Value des Analyse Tools wird in der Datenbank festgehalten
* Es kann festgelegt werden, ob der System Output des Analyse Tools in eine spezielle Output Datei geschrieben werden oder in der Datenbank gespeichert werden soll.
* Eine Logrotation verhindert allzu grosse Logdateien. Nach "n" definierten Loop Schritten wird eine neue Logdatei begonnen.

### Datenmodell

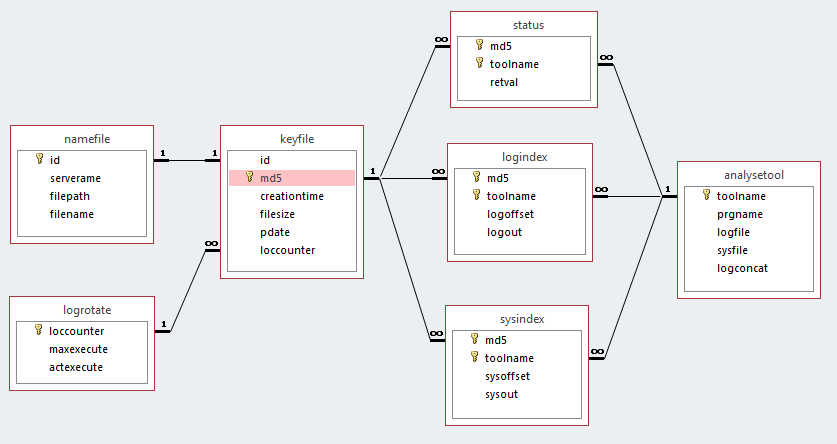
Die Tabellen *keyfile* und *namefile* enthalten die den primären Verzeichnisscan, also die Namen aller Dateien mit Dateigrösse und *Creation Time* sowie sie aus dem Lesen der Verzeichnisstrukturen erstellt werden können. Der MD5 Schlüssel wird erst beim eigentlichen Analysedurchgang angelegt, weil dafür die Dateien ebenfalls vollständig gelesen werden müssen.

Zum Ausführen der Analysemodule werden die notwendigen Informationen aus der Tabelle *analysetool* ausgelesen, d.i Programmname und Pfad, Logdatei, Datei bzw. BLOB für den System Output.

Die Tabelle *status* hält den Exit Status des Analyseprogramms fest.

Die Tabellen *logindex* und *outindex* speichern den Offset in die jeweilige Log oder Output Datei für die eben analysierte Datei.

Auf GitHub <https://github.com/KOST-CECO/TiffAnalyseProject/tree/master/createDB> sind die notwendigen SQL-loader scripts für die Datenbank abgelegt (sowohl für mySQL wie SQLite3).

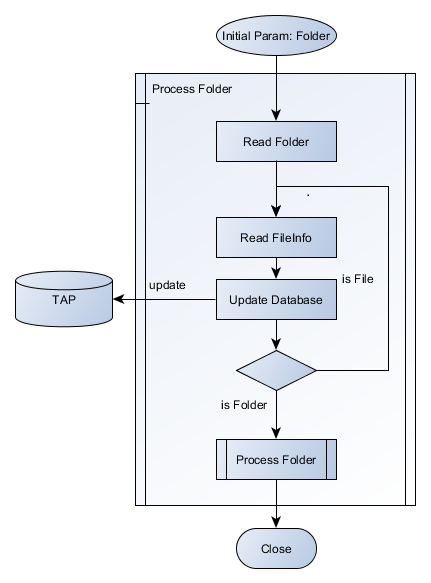


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **tablename** | **name** | **type** | **nullable** | **description** |
| **analysetool** | toolname | CHARACTER VARYING(255) | true | Name des registrierten Analyseprograms in Kurzform |
|  | prgname | CHARACTER VARYING(255) | true | Relativer Pfad und Dateiname zum Analyseprograms |
|  | logfile | CHARACTER VARYING(255) | true | Relativer Pfad und Dateiname der mit diesem Analyseprograms verbunden Logdatei |
|  | logout2file | BOOLEAN | false | Ja bedeutet, dass Logfile in "logfile" geschrieben wird sonst wird in BLOB "logout" gspeichert |
|  | sysfile | CHARACTER VARYING(255) | true | Relativer Pfad und Dateiname der mit diesem Analyseprograms verbunden Ausgabedatei |
|  | sysout2file | BOOLEAN | false | Ja bedeutet, dass die Programmausgabe in die Ausgabedatei "sysfile" geschrieben wird sonst wird in BLOB "sysout" gspeichert |
|  | logconcat | BOOLEAN | false | Ja bedeutet, dass das Programm die Logdatei fortsetzen kann, Nein es wir jeweils ein neues Log geschreiben und vom LOOP Programm an "logfile" angehängt |
|  |  |  |  |  |
| **keyfile** | id | INTEGER | true | Referenz |
|  | md5 | CHARACTER VARYING(16) | true | MD5 Hashwert |
|  | creationtime | TIMESTAMP(0) | true | Entstehungszeitpunkt der Datei laut Dateisystem |
|  | pdate | TIMESTAMP(0) | true | Zeitpunkt und Flag für den Abschluss der gesamten Analyse |
|  | loccounter | INTEGER | true | Zähler für "logfile" bzw. "sysfile" beginnend mit Eins |
|  |  |  |  |  |
| **logindex** | md5 | CHARACTER VARYING(255) | true | MD5 Schlüssel der TIFF Datei |
|  | toolname | CHARACTER VARYING(255) | true | Kurzname des Tools |
|  | logoffset | INTEGER | true | Offset in die Ausgabedatei logfile |
|  | logout | CHARACTER LARGE OBJECT | true | vollständige System Out Ausgabe des Analysetools |
|  |  |  |  |  |
| **logrotate** | loccounter | INTEGER | true | Zähler für "logfile" bzw. "sysfile" beginnend mit Eins |
|  | maxexecute | INTEGER | true | Maximal Verarbeitungsschritte pro "logfile" bzw. "sysfile" |
|  | actexecute | INTEGER | true | Aktueller Verarbeitungsschritt |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **namefile** | id | INTEGER | false | Referenz zu "keyfile" |
|  | serverame | CHARACTER VARYING(255) | true | Name des NAS Servers oder des zugeordneten Laufwerkbuchstabens |
|  | filepath | CHARACTER VARYING(255) | true | Dateipfad |
|  | filename | CHARACTER VARYING(255) | true | Dateiname mit Dateiextension |
|  |  |  |  |  |
| **status** | md5 | CHARACTER VARYING(255) | true | MD5 Schlüssel der TIFF Datei |
|  | toolname | CHARACTER VARYING(255) | true | Name des registrierten Analyseprograms in Kurzform |
|  | retval | CHARACTER VARYING(255) | true | Rückgabe Wert des Tools (Exit Status 0 = erfolgreicher Abschluss) http://www.hiteksoftware.com/knowledge/articles/049.htm |
|  |  |  |  |  |
| **sysindex** | md5 | CHARACTER VARYING(255) | true | MD5 Schlüssel der TIFF Datei |
|  | toolname | CHARACTER VARYING(255) | true | Kurzname des Tools |
|  | sysoffset | INTEGER | true | Offset in die Ausgabedatei outfile |
|  | sysout | CHARACTER LARGE OBJECT | true | vollständige System Out Ausgabe des Analysetools |

### Skript Programmierung

Wie schon oben beschrieben gliedert sich die Script Programmierung in ein "Initial Loop" Programm, das sämtliche TIFF Dateien liesst und einen Verarbeitungsindex anlegt. Das "Initial Loop" Programm soll mit unterschiedlichen und mehreren Foldern als Start Parameter umgehen können.



# Analysemodule

Folgende Analysemodule sind im Augenblick vorgesehen, weiter können nach Bedarf dazu kommen:

* [ MD5 (integriert in den Loop) ]  
  Das Berechnen des MD5 Schlüssels gehört nicht zu den Analysemodulen, wird aber vor der Analyse durchgeführt.
* Simple Formaterkennung mit *files* <http://gnuwin32.sourceforge.net/packages/file.htm>   
  Mit der Formaterkennung wird werden falsch gelabelte Dateien erkannt.
* Validierung mit JHOVE <http://jhove.openpreservation.org/>   
  Die JHOVE Validierung ermittelt die grundlegende Struktur der TIFF Datei. Wichtig sind hier *Status* und *InfoMessage*.
* DPF-Manager <http://www.preforma-project.eu/dpf-manager.html>   
  Alternative zu JOHVE aus dem Performa Projekt.
* checkit\_tiff: *a conformance checker for baseline TIFFs* der Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden  
  <https://github.com/SLUB-digitalpreservation/checkit_tiff>
* TIFF Tag Extraktion vom DHLAB (= vollständige Tagextraktion)  
  Das C++ Programm extrahiert alle TIFF Tags in eine CSV Tabelle (TIFF Tag, Datentyp und Wert)
* EXIF Extraktion mit ExifTool <http://owl.phy.queensu.ca/~phil/exiftool/>   
  Eingebettete EXIF und XMP Metadaten werden extrahiert.
* Für jede Datei wird ein sehr kleines *Thumbnail* generiert und in eine gemeinsame Log Datei geschrieben. In diesem Schritt wird erst die Payload oder Bitmap der TIFF Datei untersucht. Eine korrekte Konvertierung belegt die korrekte Implementierung von Komprimierung und Farbraum.

Folgendes soll gelten:  
Die Reihenfolge der Module ist beliebig, ausser das MD5 als erstes ermittelt werden muss. Alle Schritte werden ausgeführt, auch wenn bei einem Modul ein Fehler auftritt.

# Auswertung

Die Analysemodule werden ohne weiter Auswertung des Log- oder Systemausgabe ausgeführt. Der einzige Hinweis über das erfolgreiche Ausführen ist der Rückgabewert oder Exit Status (=retval) in der Statustabelle (=status).

Die Auswertung erfolgt vollständig offline, entweder im Archiv oder ausgelagert. Der Vorteil dieses Vorgehens ist der, dass verschiedene Auswertungen möglich sind oder dass die Auswertungsmethode auch wären der Arbeit noch verändert werden kann, etwas was in der Analysephase nicht möglich ist.

# Testdaten

Öffentlich zugängliche Testdaten:

<https://github.com/EasyinnovaSL/DPFManager/tree/develop/src/test/resources>

<https://github.com/openpreserve/jhove/tree/junit_tests/examples/tiff>

<http://sipi.usc.edu/database/database.php?volume=misc&image=7#top>