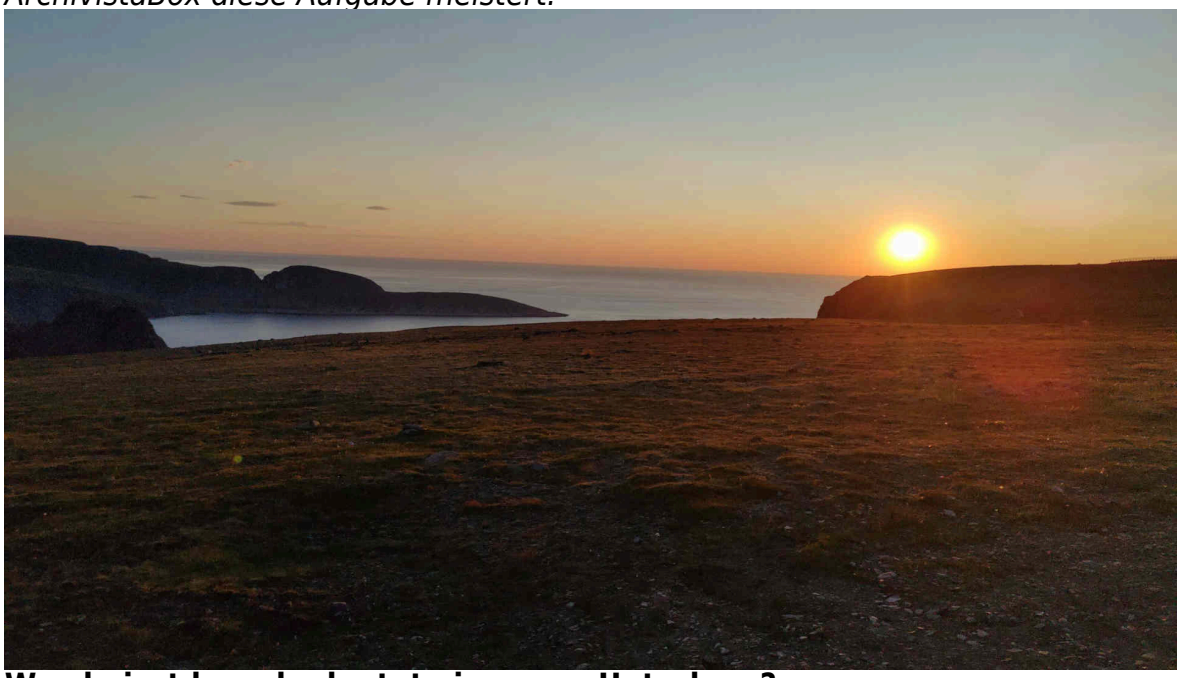
ArchivistaBox: Neues Fundament bis 2029

Egg, 20. Dezember 2024: *Uff, dies war mal wieder knapp. Längst geplant, und doch erst jetzt bereit. Das neue Fundament für die ArchivistaBox steht ab heute zur Verfügung. Für die Anwender/innen ändert sich wenig, im Support ebenfalls nicht viel mehr. Und doch ist das Update zentral, nur dank gepflegtem Unterbau laufen unsere ArchivistaBoxen über viele Jahre, gar Jahrzehnte, stabil, einfach und sicher im harten Businessalltag. In diesem Blog geht es nicht primär um das neue Release, sondern um die Komplexität digitaler Daten, und wie die ArchivistaBox diese Aufgabe meistert.*



Was heisst bzw. bedeutet ein neuer Unterbau ?

Bei der ArchivistaBox wird die Lösung aus einem Guss geliefert. Dies bedeutet, zusammen mit der Hardware wird die gesamte benötigte Software ausgeliefert, d.h. bei der die ArchivistaBox wird das Betriebssystem zusammen mit den Applikationen ausgeliefert.

Beim alten Release waren dies entweder Kernel 5.4 oder 5.10 auf der Basis von Devuan Beowulf (bzw. Debian Buster), beim neuen Unterbau wird Kernel 6.10 auf der Basis von Devuan Daedalus (bzw. Debian Bookworm) ausgeliefert. Ab Version 2024/XII werden neue Versionen ausschliesslich auf der aktuellen Basis veröffentlicht; dies im Einklang zur **Veröffentlichung im Jahre 2019.** Der Supportzeitraum für das neue Master-Release dauert bis Ende 2029.

Diese technischen Informationen seien hier der Form halber angeführt, das zentrale Thema sollen sie hier nicht sein. Bei der Realisierung des aktuellen Fundamentes stellte sich die Frage, was zentral ist. Nach Rücksprache mit vielen Kunden reifte der Entschluss, es braucht keine bahnbrechenden Neuerungen, es braucht einfach Kontinuität.

2024/XII: Stabilität und Masse

So werden mit Version 2024/XII die Scanner-Treiber neu implementiert, ebenso wurde eine einfache Remote-Lösung für alle Kunden realisiert. Einen recht hohen Aufwand bescherte der Wechsel von PHP 7.x auf 8.x, die Migration der Datenbank auf MariaDB 10.x und auch das Upgrade auf KVM/Qemu 7.x war nicht ganz trivial. Es erschien uns von daher wichtiger, dafür zu sorgen, dass alle Komponenten einfach so weiterlaufen wie sie bisher liefen.

Als Ende 2019 das letzte grosse Master-Release erschien, ging es darum, dass die ArchivistaBox multimediale Dateien verarbeiten kann. In den letzten fünf Jahren wurden diese Möglichkeiten derart erweitert, dass rückblickend angemerkt werden muss, etwas einzuführen ist die eine Sache. Es so zum Leben

zu erwecken, dass es dem harten Business-Alltag genügt, eine andere “Hausnummer”.

In den letzten Tagen wurden zum Test über 6500 Stunden Video-Material ins Archiv “geworfen”. Für das Verarbeiten der weit über 4000 Dateien benötigte die ArchivistaBox um die 5 Stunden. Dabei wurden über 760’000 Seiten neu in ArchivistaDMS angelegt. Die neu erfasste Datenmenge lag dabei bei ca. 2.2 TByte an Daten.
Bei der Verarbeitung ist kein einziger Fehler aufgetreten. Es sei hier angeführt, 2019 wäre die damalige ArchivistaBox dafür wohl noch nicht ganz bereit gewesen. Aktuell können selbst mit einer ArchivistaBox Matterhorn Archive mit mehreren Dutzend TByte bewältigt werden, von den Modellen K2 und Everest ganz zu schweigen. Mit den letzteren beiden ArchivistaBox-Modellen liessen sich Archive über Hunderte von TByte bewerkstelligen.
Und dabei werden die ArchivistaBox-Systeme auf handelsüblicher Hardware ausgeliefert, es braucht weder eine Server-Infrastruktur noch das entsprechende Know-How dazu. Damit dies möglich ist, braucht es viel Liebe zum Handwerk. Gerade bei Video-Dateien liegt der “Teufel” im Detail. Um zu erklären, wie komplex die Materie ist, sei etwas weiter ausgeholt.

Am Anfang stand das gescannte Bild

Als die Archivista-Lösung 1998 auf dem Markt erschien, ging es primär darum, gedruckte Belege über Scanner zu digitalisieren bzw. digitale Bilder ins Archiv aufzunehmen. Bereits digital erfasste Daten wurden primär als Office-Dateien und als ASCII-Textdateien gespeichert. Office-Dateien mussten damals über virtuelle PDF-Druckertreiber archiviert werden, bei den Text-Dateien besteht bis heute die Problematik, dass im Prinzip niemand weiss, mit welchem Zeichensatz die Daten gespeichert wurden.

Letztlich ging bzw. geht es noch immer darum, von Informationen Abbilder zu erhalten. ArchivistaDMS unterscheidet sich hier bis heute von Mittbewerbern, indem nicht einfach nur Quelldateien (z.B. eine Word-Datei) aufbewahrt werden, vielmehr werden die Daten virtuell “abfotografiert”. Damit ist eine visuelle Sichtung der Daten auch Jahrzehnte später problemlos möglich, selbst wenn es z.B. nicht mehr möglich sein sollte, die Quelldatei selber zu öffnen.

XML und Dokumenten-Container

Jedoch, ein Office-Dokument aus dem Jahre 2000 entspricht nicht in Ansätzen dem Aufbau aus aktueller Zeit. Wurde früher der gesamte Inhalt in einer Datei gespeichert, so werden aktuell einzelne Happen gespeichert. Um dies zu verdeutlichen, sei zum “Hallo Word”-Beispiel gegriffen. Diese 12 Zeichen (Hallo World) benötigen in den Formaten (Doc-Word, Docx-Word, PDF und Tif-Format mit 300 Pixeln/Zoll) den folgenden Speicherplatz:

halloworld.doc => 9216 Zeichen
halloworld.docx => 4206 Zeichen
halloworld.pdf => 7089 Zeichen
halloworld.tif => 2465 Zeichen

Daraus lässt sich herleiten, dass die Bild-Kopie (Tif-Datei) bis heute am effizienteste Form (rein vom Speicherplatz her gesehen) darstellt. Die PDF-Datei benötigt um den Faktor drei mehr Speicher, ohne dass das Problem der Lesbarkeit über einen langen Zeithorizont gelöst wäre (z.B. Einbettung der Schriften).

Beim docx-Format ist anzufügen, dass es sich hier um XML-Daten in komprimierter Form handelt. Wird die entsprechende Word-Datei ausgepackt, so erscheinen die folgenden Dateien:

./docProps
./docProps/core.xml
./docProps/app.xml
./[Content_Types].xml
./_rels
./_rels/.rels
./word
./word/fontTable.xml
./word/document.xml
./word/_rels
./word/_rels/document.xml.rels

./word/styles.xml
./word/settings.xml
Und dies alles, für “satte” 12 Zeichen. Ausgepackt werden dafür auf der Festplatte um die 56’000 Zeichen (Bytes) benötigt. Soviel zur Effizienz heutiger Computerprogramme. Unabhängig davon ist es aber doch so, je mehr Daten bei einem Dateiformat anfallen, desto mehr Komplexität ergibt sich daraus, um diese Informationen langfristig und sicher am Leben zu erhalten.

Video-Dateien als Komplexitätswalze
Der kleine Ausflug in gescannte Daten und die Office-Formate dient hier als Einstieg zu den Video-Dateien. Im Prinzip könnte ja angenommen werden, es handle sich bei Video-Dateien um Abfolgen von Bildern, angereichert durch etwas Ton.
Leider ist die Sachlage bei den Video-Dateien aber weit komplexer. Nachfolgend soll es hier primär um das MP4-Format (im engeren Sinn um H264) gehen. Auch hier handelt es sich (analog zu den Office-Dateien) um eine Spezifikation, bei der mehrere Brocken zusammen abgelegt werden können. Rein von der Spezifikation her können darin auch mehrere Video-Daten miteinander gesichert werden. Allerdings kann bei so erstellten Video-Dateien von den meisten Playern nur die erste Spur abgespielt werden.
Dagegen können viele verschiedene Audio-Quellen vorhanden sein, in unterschiedlicher Qualität, wobei es meist darum geht, mehrere Sprachen in der Datei aufzunehmen. Dazu kommen Untertitel-Dateien, die entweder als Bitmap oder (aktuell) als Text-Dateien in den MP4-Dateien integriert sind.

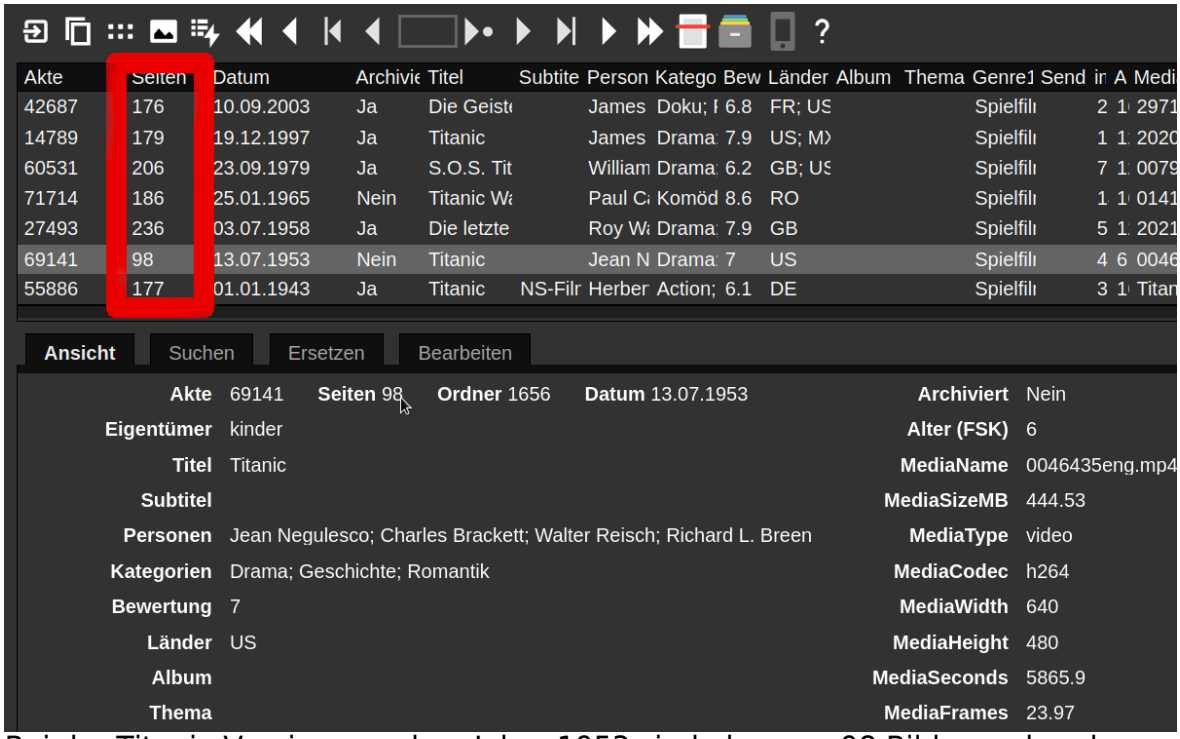
Der Titanic-Dampfer versinkt mitten in der Datei

In den letzten Jahren wurden viele Tausende von Dateien zu Testzwecken verarbeitet. Dabei musste festgestellt werden, dass es auch Video-Dateien gibt, bei welchen z.B. die Audiospur erst eine Sekunden ab Beginn der Video-Spur vorhanden ist. Gute Video-Abspieler kommen damit klar, bei Firefox z.B. aber funktionierte es nicht (die Ton- und Video-Spur stimmten dann nicht).

Bei den Untertiteln musste verschiedentlich festgestellt werden, dass nicht korrekte Daten vorhanden waren. So konnte z.B. beim Titanic-Film aus dem Jahre 1953 (lässt sich mühelos im Netz finden) festgestellt werden, dass Untertitel der Version aus dem Jahre 1998 mitgeliefert wurden. Die 1998-er-Version dauert satte 3 Stunden und 14 Minuten. 1953 dauerte der “Kampf” nur bis zur Halbzeit (1 Stunde 37 Minuten).

Der Hauptkandidat für das Verarbeiten von MP4-Dateien (ffmpeg) sieht darin kein Problem. Die Untertitel-Datei aus dem Jahre 1998 mit Zeitangaben über 3 Stunden kann problemlos in der Video-Datei des Jahres 1953 abgelegt werden. In der Praxis versinkt der Dampfer in der Version von 1953 dann “pünktlich” bei 1 Stunde und 37 Minuten, auch wenn einige Player zum Abspielen bis weit über 3 Stunden einladen.

Entdeckt wurde dies zufälligerweise bei der Ablage des Films im Privatarchiv des Geschäftsführers. Dies deshalb, weil ArchivistaDMS bei den Video-Dateien Vorschaubilder erstellt. Natürlich können dabei nicht alle Bilder einzeln abgelegt werden, dies würde ja zu unhaltbar grossen Datenbeständen führen. Vielmehr ist es so, dass (ausser bei sehr kurzen Video-Dateien) etwa zwischen 150 bis 300 Vorschaubilder erstellt werden.



Bei der Titanic-Version aus dem Jahre 1953 sind aber nur 98 Bilder vorhanden. Beim manuellen Überprüfen musste festgestellt werden, dass der Film die falschen Untertitel hatte (jene aus der über dreistündigen Version von 1998).

MP4Test: Damit die Titanic nicht mehr abschifft…

Um die Integrität von Video-Dateien zu überprüfen, kommt meist das Konsolenprogramm 'ffmpeg' zum Einsatz. Bei **superuser.com** wird z.B. empfohlen, entsprechende Videos mit folgendem Befehl zu überprüfen:

```
ffmpeg -v error -i titanic1953.mp4 -f null - 2>error.log
```

Allerdings generiert der obige Test bei besagter Datei keine Fehler, d.h. die Datei wird für gut befunden. Letztlich ist es wohl so, nur wenn jedes einzelne Bild aus dem Video extrahiert würde, erst dann könnte festgestellt werden, ob Fehler vorliegen. Ein derartiger Test erforderte aber sehr viel Rechenzeit. Für eine grosse Anzahl von Dateien wäre ein solcher Test kaum praktikabel. Das hier vorgestellte Helferlein mp4test ist darauf ausgelegt, einzelne Dateien oder ganze Ordner zu überprüfen:

```
/usr/bin/mp4test file|pathorfilein -- Program to check mp4 files for errors
(c) 20.12.2024 by Archivista GmbH, Urs Pfister, https://archivista.ch
Licence: GPLv2, Deps: ffmpeg,ffprobe,grep,pgrep (tested under Linux)
Status messages go to console and log file (mp4test.log)
```

Der Sourcecode kann **hier bezogen werden;** er befindet sich aber auch auf jeder AVMultimedia/ArchivistaBox-Version ab 2024/XII. Mit diesem Programm wird eine Schattenkopie erstellt, wobei (um den Vorgang entsprechend zu beschleunigen) jeweils nur 1 Bild pro Sekunde extrahiert wird.
Dadurch werden pro Stunde Material noch etwa 15 bis 20 Sekunden benötigt. Dadurch, dass das Programm mit mehreren Prozessen arbeitet, können pro Minute bei einem Prozessor mit 8 Kernen (CPUs) in etwa 10 Stunden recht zuverlässig auf Fehler überprüft werden. Anhand der Titanic-Datei von 1953 kann gezeigt werden, wie mp4test arbeitet. Auf der ArchivistaBox einfach im entsprechenden Ordner die gewünschte Datei aufrufen:

```
mp4test titanic.mp4
Wait for titanic.done (5)
```

Wait for titanic.done (10)
Wait for titanic.done (15)
Wait for titanic.done (20)
ERROR: titanic.mp4 - time shoud be: 03:12:57.21, but is 01:37:46.00
Unschwer zu erkennen ist, dass für den Test gute 20 Sekunden notwendig waren. Bei einem Fehler wird eine entsprechende Fehlermeldung generiert. Bei der korrigierten Version ergibt sich die folgende Meldung:
Wait for titanic.done (5)
Wait for titanic.done (10)
Wait for titanic.done (15)
Wait for titanic.done (20)
File titanic.mp4 has 01:37:45.92, check has 01:37:46.00, is ok
Da nur jede Sekunde ein Bild getestet wird, stimmt am Ende die Zeit nicht zu 100%. Je nach der Anzahl der Bilder pro Sekunde bleiben minimale Differenzen übrig. Fehler werden immer dann generiert, wenn die Inhaltslänge um mehr als 2 Sekunden von der neu erstellten Schattenkopie abweicht.
Mit Angabe eines Verzeichnisses kann eine beliebige Anzahl von Video-Dateien in einem Rutsch überprüft werden. Dabei können bei 8 Kernen wohl ca. 4’000 bis 6’000 Videos bzw. ca. 10’000 Stunden pro Tag überprüft werden.
Das kleine Hilfsprogramm mp4test zeigt gut auf, wo die Tücken bei digitalen Dateien am Beispiel von Video-Dateien liegen und welcher Aufwand bei der ArchivistaBox betrieben wird, um Daten langfristig verfügbar zu halten. In diesem Sinne, Enjoy!

Betriebsferien 23. Dezember 2024 bis 3. Januar 2025
Die “offiziellen” Feiertag dauern dieses Jahr vom **Montag, 23. Dezember 2025 bis und mit Freitag 3. Januar 2026.** Im Unterschied zu anderen Jahren ist es aber so, dass bedingt durch die aufwändigen Arbeiten beim neuen Release 2024/XII noch viele Arbeiten (z.B. Versand der Rechnungen) zu erledigen sind. Diese Arbeiten erfolgen dieses Jahr wohl zwischen Weihnachten und Neujahr. Kunden mit Wartungsverträgen erhalten an den Werktagen (23.12, 27.12, 30.12 und 3.1) selbstverständlich Support. Zum Abschluss möchte ich mich an dieser Stelle bei unserer Kundschaft für die langjährige Treue bedanken. Mit grosser Freude stehe ich im Jahr 2025 für ihre Anliegen zur Verfügung, ebenso wie die ArchivistaBox (gerade mit dem neuen Master-Release) selbstverständlich

weiterentwickelt wird. In diesem Sinne, besinnliche Weihnachtstage und einen guten Rutsch
Urs Pfister, Archivista GmbH

P.S: Die Bilder in diesem Blog stammen aus Schweden, Finnland und Norwegen. Das obige Bild des Geschäftsführers stammt vom 8.8.24, nachdem er in 19 Tagen von Italien bis zum Nordkapp mehr als 4200 Kilometer mit dem Velo bewältigt hat (an besagtem Tag waren es knapp 300 Kilometer). Dazu gibt es einen Dokumentarfilm, mehr dazu unter ***Dall'Italia til Nordkapp.*** *Alle Kunden erhalten für den Film auf Anfrage ein kostenfreies Online-Ticket.*

Facebook

Twitter