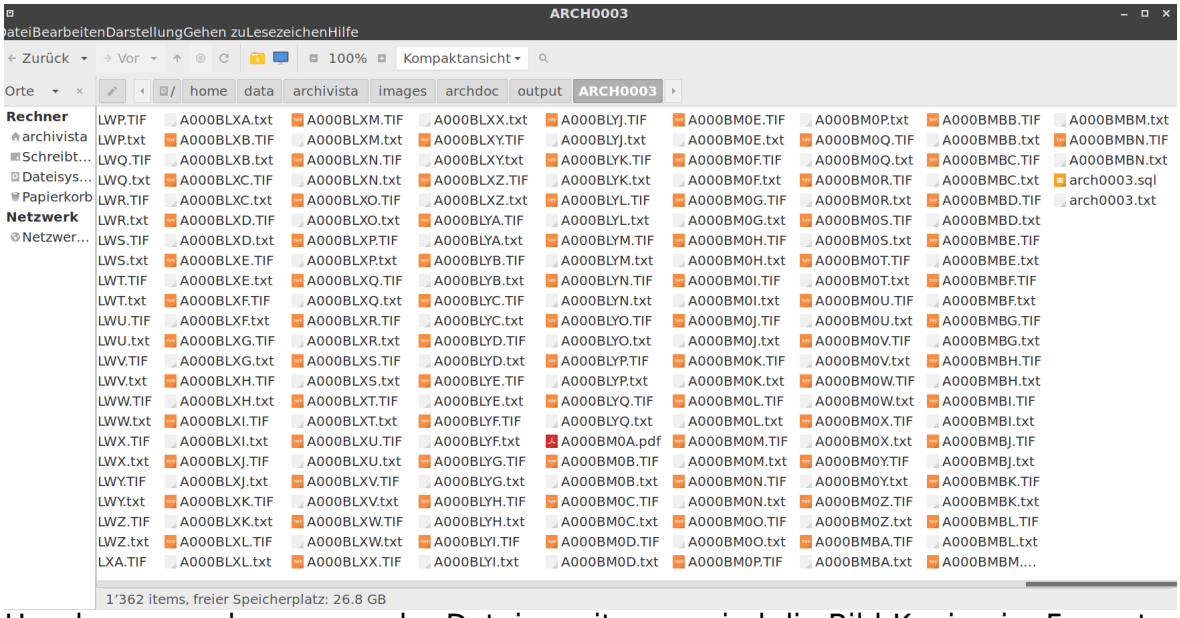
30 Jahre Datenintegrität erweitert
Egg, 4. Juli 2024: *Das Produkt Archivista gibt es seit 1998, die ArchivistaBox seit 2006. In der Informatiklandschaft sind dies ordentliche Zeiträume, in der Geschichte der Erde (um es pathetisch zu sagen) ist es ein Klacks. Und doch ist es so, ohne Daten sind wir heutzutage ein Nichts. Die ArchivistaBox-Lösung bietet dabei ein einzigartiges Konzept für die langfristige Datensicherheit an. Mit der neuen Version 2024/VII wird diese nochmals erweitert und darum geht es in diesem Blog.*

Fünf Säulen der Datensicherheit
Manche denken, ein DMS-System sei faktisch einzig ein relativ geordneter Haufen von Dateien bzw. dies könne auch über das Betriebssystem erledigt werden. Nüchtern betrachtet ist dies deutlich zu kurz gegriffen, ein DMS umfasst (unter anderem) automatisierte Abläufe und ein ebensolches Berechtigungskonzept.
Die ArchivistaBox bietet aber nicht nur ausgiebige Möglichkeiten beim Management der Daten an, sondern darüber hinaus (und insbesondere) ausgiebige Sicherungskonzepte für die Daten. Nachfolgend wird das Konzept vorgestellt, um am Ende auf die Neuerungen der Version 2024/VII einzugehen.
Stufe 1: Erstellen von Bild-Daten
Damit die visuelle Lesbarkeit der Daten gewährleistet ist, werden von sämtlichen Dokumenten, die in ArchivistaDMS aufgenommen werden, Bild-Daten erstellt. Die Daten werden dabei virtuell abfotografiert. Dieses «Abfotografieren» erfolgt mit hoher Geschwindigkeit. Auf einen **MediaVM-Server Everest** können pro Tag mehrere Millionen Seiten verarbeitet werden, auf einer **ArchivistaBox-Dolder** könnten rein rechnerisch noch immer pro Sekunden einige Seiten verarbeitet werden. Ziel dieses ersten Schrittes ist es, die Lesbarkeit aller Daten (z.B. Office-Dateien oder Mails) ohne externe Plugins in ArchivistaDMS zu gewährleisten.
Stufe 2: Log-Dateien der Datenbank
Wird ein Dokument in die ArchivistaBox aufgenommen, so erfolgt die Speicherung in der Datenbank. Dabei wird der Inhalt gleichzeitig in Log-Dateien zusätzlich gesichert. Ein Dokument, das (später) gelöscht wird, befindet sich zu diesem Zeitpunkt zwar nicht mehr aktiv (zugreifbar) in der Datenbank. Solange die Log-Dateien der Datenbank nicht gelöscht werden, sind die Daten dort jedoch noch verfügbar.

Stufe 3: Redundanz auf Stufe Hardware
Ab Ausbaustufe Titlis werden bei der ArchivistaBox immer zwei Systeme aufgebaut. Ganz egal, ob es sich dabei um physikalische Boxen oder virtualisierte Instanzen handelt. Die erste Box ist dafür zuständig, die Daten zu erfassen, auf dem zweiten Gerät erfolgt eine Spiegelung der ersten Box. Bei einem Ausfall der Hardware sind tagesaktuelle Änderungen solange noch verfügbar, wie nicht beide Systeme gleichzeitig ausfallen.
Stufe 4: Sicherung auf externe Datenträger
Mit der (periodischen) Datensicherung werden die Daten auf externe Datenträger (es kann auch ein externer Computer (selbst die Cloud) sein. Die so gesicherten Daten (Backup) lassen sich bei Bedarf wieder zurückspielen. Angemerkt sei, dass dieser Prozess (Restore) einzig auf dem ArchivistaBox-Desktop über den Menupunkt ‚ArchivistaSetup' möglich ist. Dies hat den Vorteil, dass über das Web-Interface niemals die aktuellen Daten überschrieben werden können.
Stufe 5: Langfristige Sicherung im ISO-9660-Format
Mit der Sicherung auf externe Datenträger / Rechner begnügen sich die allermeisten Lösungen, sei dies bei DMS-Systemen oder ganz generell in der Informatik. Dabei besteht die Problematik, dass so gesicherte Daten nicht vor Veränderungen gesichert sind. Selbst wenn Prüfziffern zu den Daten erstellt werden, so kann bei einer Manipulation bzw. dem Verlust der Daten lediglich festgestellt werden, dass die korrekten Informationen fehlen.
Hier setzt die Langzeitsicherung der ArchivistaBox ein. Entweder von Hand oder in periodischen Zeitabständen wird ein Archivierungsprozess angeworfen. Dabei werden ISO-Dateien erstellt, auf denen sämtliche Daten unabhängig in einer einfachen Ordner-Struktur von der aktuellen Version der ArchivistaBox gesichert werden. Darin enthalten sind die Bild- und die Struktur-Daten.
Erweiterte Struktur bei Langzeitkopie (ab Version 2024/VII)
Nachfolgend geht es um die Struktur der Langzeitkopie bei der ArchivistaBox (Stufe 5). Zunächst, auf dieser Ebene, und dies dürfte einzigartig in der Landschaft der DMS-Systeme sein, gibt es 30 Jahre Garantie für die Datenintegrität. D.h. die Lesbarkeit der Daten wir für minimal drei Jahrzehnte auf jedem handelsüblichen Computer gewährt, und zwar unabhängig von einer bestimmten Version der ArchivistaBox, ja gar ganz ohne die ArchivistaBox an sich. Dazu wird auf die ISO 9660-Norm zurückgegriffen.
Kern dieser Norm ist, dass Dateinamen auf 8 Buchstaben (plus 3 für die Erweiterung) beschränkt sind. Ebenso gibt es keine Umlaute und die Anzahl der Ordnerebenen ist auch limitiert. Es mag 2024 anachronistisch erscheinen, derart kurze Dateinamen zu verwenden. Aber ganz ehrlich gesagt, was bringen gesicherte Daten, wenn bereits auf Stufe Dateiname später z.B. die Umlaute nicht korrekt dargestellt werden können?
Die nachfolgende Abbildung zeigt einen Teil eines Ordner-Abbildes (konkret Ordner 3).



Unschwer zu erkennen an der Dateierweiterung sind die Bild-Kopien im Format TIF, PNG oder JPG. Nur, welchen Kontext nehmen diese Dateien ein?
Die Zuordnung einer Datei zu einer Akte bzw. Seite erfolgt über die Namensdefinition der ArchivistaBox (siehe **ArchivistaBox-Handbuch** bzw. dort **Datenstrukturen).** Ab Version 2024/VII gibt es neu ein Hilfsprogramm, dass die Akte und Seite aus einer archivierten Datei berechnet. Das entsprechende Programm **getdocpage.pl** befindet sich im Ordner **/home/cvs/archivista/jobs,** nachfolgend wird der Code hier publiziert:

```
#!/usr/bin/perl
# Program to calculate Doc and Page from archived files
(ArchivistaBox)
# v0.1 (c) 2024-07-02 by Archivista GmbH, Urs Pfister
use strict;
my $name = shift; # (base) file name
my $short = shift; # 0=full message, 1=short message
$name = uc($name);
exitgo("filename to print out Doc and Page of a document")
if $name eq "";
$name = substr($name,0,8) if length($name)==12;
if (length($name)!=8) {
  exitgo("filename $name must have 12 chars
[A000XXYY.TIF|JPG|PNG]ZIP]");
}
my $doc = calcNumber(substr($name,0,6));
my $page = calcNumber(substr($name,6,2));
if ($short==0) {
  print "$name => Doc: $doc / Page: $page\n";
} else {
  print "$doc-$page\n";
}

sub calcNumber {
  my ($part) = @_;
  my $lang = length($part);
  my $number=0;
  my $number1=0;
  my $c2=0;
  for (my $c=$lang;$c>=1;$c--) {
```

```
    my $c1=$c-1;
    my $charminus=65;
    my $part1 = ord(substr($part,$c1,1))-$charminus;
    if ($part1>=0) {
      $number1 = $part1*(26**$c2);
      $number = $number + $number1;
    }
    $number1=0;
    $c2++;
  }
  return $number;
}

sub exitgo {
  my ($msg) = @_;
  print "$0 $msg\n";
  exit 1;
}
```

Um von einer archivierten Seite die entsprechende Akte bzw. Seite zu erhalten, kann das Programm unter Angabe der Datei aufgerufen werden:

perl getdocpage.pl A000BMBN.TIF

Daraus resultiert die Anzeige:

A000BMBN => Doc: 38 / Page: 39

Bislang wurde bei der ArchivistaBox der entsprechende SQL-Datenstrom (Structured Query Language) gesichert. Dies entspricht beim Ordner 3 der **Datei arch0003.sql.** Darin finden sich sämtliche Informationen der Datentabelle **archiv** als SQL-Befehle. Auch wenn SQL an sich ein Standard ist, interpretieren Datenbank-Systeme SQL nicht immer gleich bzw. korrekt. Daher findet sich **ab Version 2024/VII zusätzlich eine ANSI-Datei mit den Strukturinformationen.** In unserem Beispiel trägt diese den Namen **arch0003.txt.**

Darin finden sich die Informationen zu den einzelnen Akten sowie auf der letzten Spalte die Namen der Bild-Dateien. Mit dieser neu erstellten Strukturdatei dürfte es noch einfacher sein, so archivierte Daten einfach einzusehen bzw. mit einem Skript automatisiert auszulesen bzw. zu verarbeiten.

Ebenfalls neu in Version 2024/VII ist, dass der dazugehörende Text einer Seite (sofern vorhanden) mit der Dateiendung ‚txt' ebenfalls vorhanden ist. Damit kann direkt in den archivierten Ordnern nach dem Text der Dokumente gesucht werden.

Angemerkt an dieser Stelle sei noch (auch wenn dies kein neues Feature der Version 2024/VII ist), dass mit der Endung ‚.pdf' allfällig vorhandene PDF-Dateien vorliegen und bei den Dateien mit der Endung ‚ZIP' finden sich die originären Dokumente (z.B. Office-Datei) in gezippter Form. Multimediale Dateien enthalten die jeweilige Dateiendung (MP3,MP4,OGG), wobei Dateien über 4 GByte (betrifft Filme) in Datenblöcken zu jeweils 512 MByte vorliegen. Letzterer Punkt ist der ISO-9660-Kompatiblität gefordert, da Dateien über 4 GByte nicht zulässig sind.

Darum sind Archivdatenträger (M-Disk) sinnvoll
Beim Archivprozess werden im ersten Schritt die entsprechend oben beschriebenen Ordner erstellt. In einem zweiten Schritt werden dazu passende ISO-Dateien erstellt, sofern genügend Daten für einen kompletten Archivdatenträger vorhanden sind.
Historisch gesehen wurden die archivierten Datenträger für lange Zeit auf **CD-R-Scheiben (C**ompact **D**isc **R**ecordable) gebrannt. Das Dateivolumen von ca. 700 MByte erscheint für heutige Verhältnisse aus der Zeit «gefallen». Immerhin, schon vor der Jahrtausendwende konnten damit ca. 10’000 Seiten pro Scheibe langfristig archiviert werden.
Später kam das **DVD-Format** (**D**igital **V**ideo **D**isc, 4.2 GByte) dazu. Bei Archivgrössen im TByte-Bereich sind sowohl CDR als DVDs kaum mehr praktikabel, da pro TByte ca. 1300 CDRs bzw. noch immer ca. 240 DVDs erstellt werden müssten. Mit dem M-Disk-Format (100 GByte) sind es noch 10 bzw. 20 Datenträger (bei zwei Kopien). Damit können selbst grosse (auch multimediale) Archive sauber auf nicht wiederbeschreibbare Datenträger ausgelagert werden. Die dazu notwendigen Informationen finden sich im **Archivista-Handbuch** unter **‚Ordner brennen‘.**
Epilog 1: Nachträgliches Erstellen der Archiv-Ordner
Auf jeder ArchivistaBox befindet sich im Ordner **/home/cvs/archivista/jobs** das Programm **extract-folder.pl,** um allfällig bereits gelöschte Archivordner neu zu erstellen. Der Aufruf (root-Benuzter) erfolgt wie folgt:
`cd /home/cvs/archivista/jobs;perl extract-folder dbname 1 10`
Bei dbname ist der Name der Datenbank anzugeben und 1 bzw. 10 entsprechenden den gewünschtem Start- bzw. Endordner.
Hinweis: *Das obige Hilfsprogramm benötigt für komplette Archive genügend Speicherplatz (minimal 50% der Festplatte muss frei sein) und kann unter Umständen viele Stunden bis Tage benötigen.*
Epilog 2: Exportieren/Importieren von Daten
Soll der gesamte Inhalt einer Datei in einen einzelnen Ordner exportiert werden, besteht dafür das Hilfsprogramm **avimportexport2.pl** (Ordner **/home/cvs/archivista/jobs).** Ohne Angabe von Parametern erfolgt die folgende Ausgabe:
`avimpexport2.pl importdb2|exportdb2 dbname dir docx-docy|NULL sql`
An erster Stelle anzugeben ist, ob ein Export (exportdb2) oder Import

(importdb2) erfolgen soll. Beim zweiten Parameter ist der Name des Archivs anzugeben. Beim dritten Wert ist der gewünschte Pfad zu bestimmen. Danach (vierte Angabe) ist der Dokumentenbereich anzugeben (z.B. 20-40). Beim Export kann alternativ beim vierten Wert auch NULL angegeben werden und als fünfter Parameter eine SQL-Bedingung (z.B. Seiten=1).
Hinweis: *Das obige Hilfsprogramm benötigt je nach Anzahl der Dokumente über genügend Speicherplatz (bis 50% der Festplatte muss frei sein) und kann unter Umständen viele Stunden bis Tage benötigen. Falls PDF-Dateien erstellt werden sollen, kann das Programm* ***exportpdf.pl*** *verwendet werden. Dabei entfällt der erste Parameter.*

Fazit: Niemand muss, aber jede/r sollte

Mit den erweiterten bzw. den neuen Möglichkeiten bei Version 2024/VII lässt sich die langfristige Datensicherheit erheblich erweitern. Natürlich darf jede/r einfach darauf vertrauen, dass Festplatten die Daten ‚fest' speichern. Nur muss sich niemand wundern, wenn auf den Festplatten die Daten dennoch ‚platt' sind, wenn es nicht so optimal läuft.

Denn eines ist klar, auch wenn der bisherige Sommer seinem Namen nicht so wirklich gerecht wurde, der nächste wirklich heisse Sommer kommt bestimmt. Abgesehen davon, Nässe überstehen Festplatten ebenfalls nicht. In diesem Sinne, einen angenehme Ferienzeit… Bleibt anzufügen, in der Zeit zwischen **19. Juli 2024 und 9. August (Sommerpause)** werden **keine neuen ArchivistaBoxen** ausgeliefert. Selbstverständlich erhalten Kunden mit Wartungsvertrag Support.