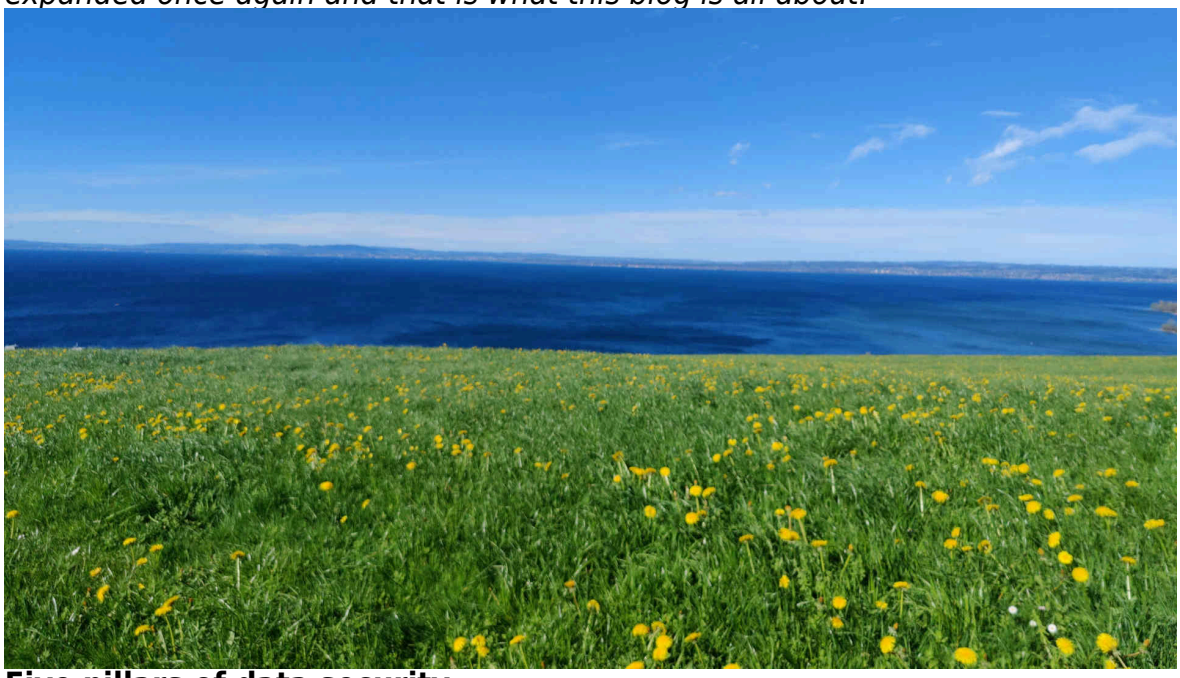
30 years of data integrity extended

Egg, July 4, 2024: *The Archivista product has been around since 1998, the ArchivistaBox since 2006. In the IT landscape, these are decent periods of time, in the history of the earth (to put it pathetically) it's a cinch. And yet, without data, we are nothing these days. The ArchivistaBox solution offers a unique concept for long-term data security. With the new version 2024/VII, this is being expanded once again and that is what this blog is all about.*



Five pillars of data security

Some people think that a DMS system is in fact just a relatively well-organized pile of files or that this can also be done via the operating system. Viewed objectively, this is clearly too short-sighted; a DMS comprises (among other things) automated processes and a similar authorization concept.
However, ArchivistaBox not only offers extensive data management options, but also (and in particular) extensive data backup concepts. The concept is presented below and the new features of version 2024/VII are discussed at the end.

Level 1: Creating image data

To ensure the visual readability of the data, image data is created for all documents that are recorded in ArchivistaDMS. The data is photographed virtually. This "photographing" takes place at high speed. Several million pages can be processed per day on an **Everest MediaVM server,** while a few pages could still be processed per second on an **ArchivistaBox Dolder** in purely mathematical terms. The aim of this first step is to ensure the readability of all data (e.g. Office files or emails) without external plugins in ArchivistaDMS.

Level 2: Log files of the database

When a document is added to the ArchivistaBox, it is saved in the database. At the same time, the content is also saved in log files. A document that is (later) deleted is no longer active (accessible) in the database at this time. However, as long as the database log files are not deleted, the data is still available there.

Level 3: Redundancy at hardware level

From the Titlis expansion level, two systems are always set up for the ArchivistaBox. Regardless of whether these are physical boxes or virtualized instances. The first box is responsible for recording the data, while the first box is mirrored on the second device. In the event of a hardware failure, daily changes are still available as long as both systems do not fail at the same time.

Level 4: Backup to external data carriers
With the (periodic) data backup, the data is backed up to external data carriers (this can also be an external computer (even the cloud). The data backed up in this way can be restored if required. It should be noted that this process is only possible on the ArchivistaBox desktop via the menu item ‘ArchivistaSetup’. This has the advantage that the current data can never be overwritten via the web interface.
Stage 5: Long-term backup in ISO-9660 format
The vast majority of solutions are content with backing up to external data carriers/computers, whether for DMS systems or in IT in general. The problem here is that data backed up in this way is not protected against manipulation. Even if check digits are created for the data, if the data is manipulated or lost, it can only be determined that the correct information is missing.
This is where the long-term backup of the ArchivistaBox comes in. An archiving process is started either manually or at periodic intervals. ISO files are created in which all data is backed up independently of the current version of the ArchivistaBox in a simple folder structure. This contains the image and structure data.
Extended structure for long-term copy (from version 2024/VII)
The following is about the structure of the long-term copy in ArchivistaBox (level 5). First of all, at this level, and this is probably unique in the landscape of DMS systems, there is a 30-year guarantee for data integrity. This means that the readability of the data is guaranteed for a minimum of three decades on any standard computer, regardless of a specific version of the ArchivistaBox or even without the ArchivistaBox itself. The ISO 9660 standard is used for this purpose. The core of this standard is that file names are limited to 8 characters (plus 3 for the extension). There are also no umlauts and the number of folder levels is also limited. It may seem anachronistic in 2024 to use such short file names. But to be honest, what is the point of saved data if, for example, the umlauts cannot be displayed correctly at the file name level?
The following illustration shows part of a folder image (specifically folder 3).

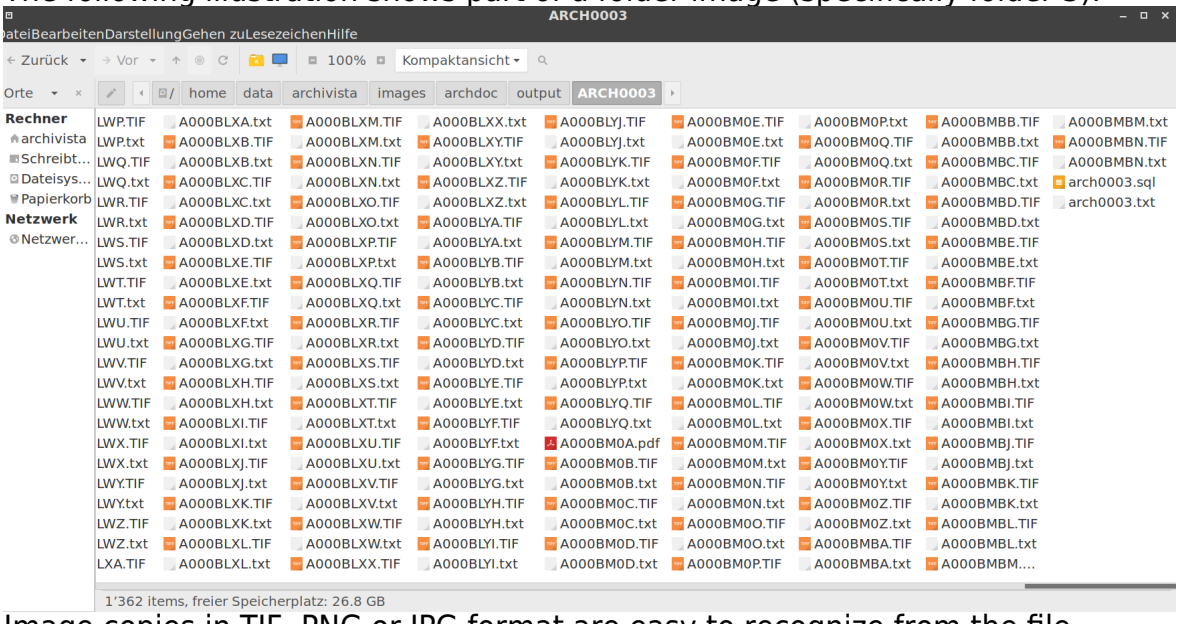


Image copies in TIF, PNG or JPG format are easy to recognize from the file extension. But what context do these files have?
A file is assigned to a record or page via the name definition of the ArchivistaBox (see **ArchivistaBox manual** or **data structures** there). As of version 2024/VII, there is a new auxiliary program that calculates the file and page from an archived file. The corresponding program **getdocpage.pl** is located in the folder **/home/cvs/archivista/jobs,** the code is published here

below:

```
#!/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;
}
```

To obtain the corresponding file or page from an archived page, the program can be called up by specifying the file:

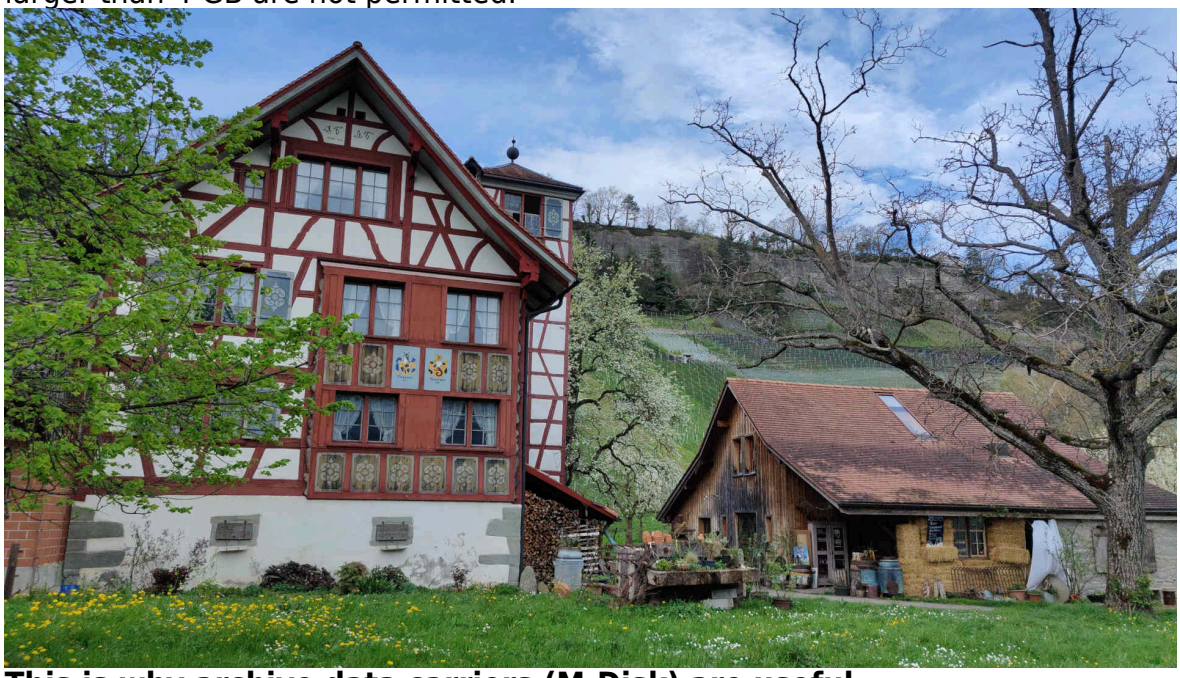
perl getdocpage.pl A000BMBN.TIF

This results in the display:

A000BMBN => Doc: 38 / Page: 39

Previously, the corresponding SQL data stream (Structured Query Language) was saved in the ArchivistaBox. This corresponds to the **arch0003.sql** file in folder 3. It contains all the information of the archiv data table as SQL

commands. Even if SQL itself is a standard, database systems do not always interpret SQL in the same way or correctly. For this reason, **as of version 2024/VII, there is also an ANSI file with the structure information.** In our example, this has the name **arch0003.txt.**
It contains the information on the individual files as well as the names of the image files in the last column. With this newly created structure file, it should be even easier to simply view archived data or to automatically read or process it with a script.
Another new feature in version 2024/VII is that the corresponding text of a page (if available) is also available with the file extension 'txt'. This means that the text of the documents can be searched for directly in the archived folders.
It should also be noted at this point (even if this is not a new feature of version 2024/VII) that any existing PDF files are available with the extension '.pdf' and the files with the extension 'ZIP' contain the original documents (e.g. Office file) in zipped form. Multimedia files contain the respective file extension (MP3, MP4, OGG), whereby files over 4 GB (concerns films) are available in data blocks of 512 MB each. The latter point is required for ISO 9660 compatibility, as files larger than 4 GB are not permitted.



This is why archive data carriers (M-Disk) are useful
In the first step of the archive process, the folders described above are created. In a second step, suitable ISO files are created, provided there is enough data for a complete archive disk.
Historically, archived data carriers were burned onto **CD-R disks** (**C**ompact **D**isc **R**ecordable) for a long time. The file volume of approx. 700 MByte seems "out of date" by today's standards. Nevertheless, even before the turn of the millennium, around 10,000 pages per disk could be archived long-term.
The **DVD format** (**D**igital **V**ideo **D**isc, 4.2 GByte) was added later. With archive sizes in the TByte range, both CDRs and DVDs are hardly practicable any more, as approx. 1300 CDRs or still approx. 240 DVDs would have to be created per TByte. With the M-Disk format (100 GByte) it is still 10 or 20 data carriers (with two copies). This means that even large (including multimedia) archives can be stored cleanly on non-rewritable data carriers. The necessary information can be found in the **Archivista manual** under **'Burn folder'.**
Epilogue 1: Subsequent creation of archive folders
The program **extract-folder.pl** is located in the folder **/home/cvs/archivista/jobs** on every ArchivistaBox in order to recreate any

archive folders that have already been deleted. The call (root user) is made as follows:
`cd /home/cvs/archivista/jobs;perl extract-folder dbname 1 10`
For dbname, enter the name of the database and 1 or 10 corresponding to the desired start or end folder.
Note: *The above utility requires sufficient storage space for complete archives (at least 50% of the hard disk must be free) and may take many hours or even days to complete.*

Epilogue 2: Exporting/importing data

If the entire content of a file is to be exported to a single folder, the utility program **avimportexport2.pl** (folder **/home/cvs/archivista/jobs)** is available for this purpose. The following output is produced if no parameters are specified:
`avimpexport2.pl importdb2|exportdb2 dbname dir docx-docy|NULL sql`
The first parameter to be specified is whether an export (exportdb2) or import (importdb2) is to take place. The name of the archive must be specified for the second parameter. The desired path must be specified for the third value. Then (fourth value) the document range must be specified (e.g. 20-40). Alternatively, NULL can be specified for the fourth value and an SQL condition (e.g. pages=1) as the fifth parameter.
Note: *Depending on the number of documents, the above utility requires sufficient storage space (up to 50% of the hard disk must be free) and may take many hours or even days to complete. If* ***PDF files*** *are to be created, the* ***exportpdf.pl*** *program can be used. In this case, the first parameter is omitted.*

Conclusion: Nobody has to, but everyone should

With the extended options (or the new options in version 2024/VII), long-term data security can be considerably enhanced. Of course, everyone can simply trust that hard disks store data ‘permanently’. But nobody should be surprised if the data on the hard disks is still ‘flat’ when things are not going so well. Because one thing is clear, even if the summer so far has not really lived up to its name, the next really hot summer is sure to come. Apart from that, hard disks don’t survive the wet either. With this in mind, have a pleasant vacation season… It remains to be added that **no new ArchivistaBoxes** will be delivered between **July 19, 2024 and August 9 (summer break).** Of course, customers with a maintenance contract will receive support.