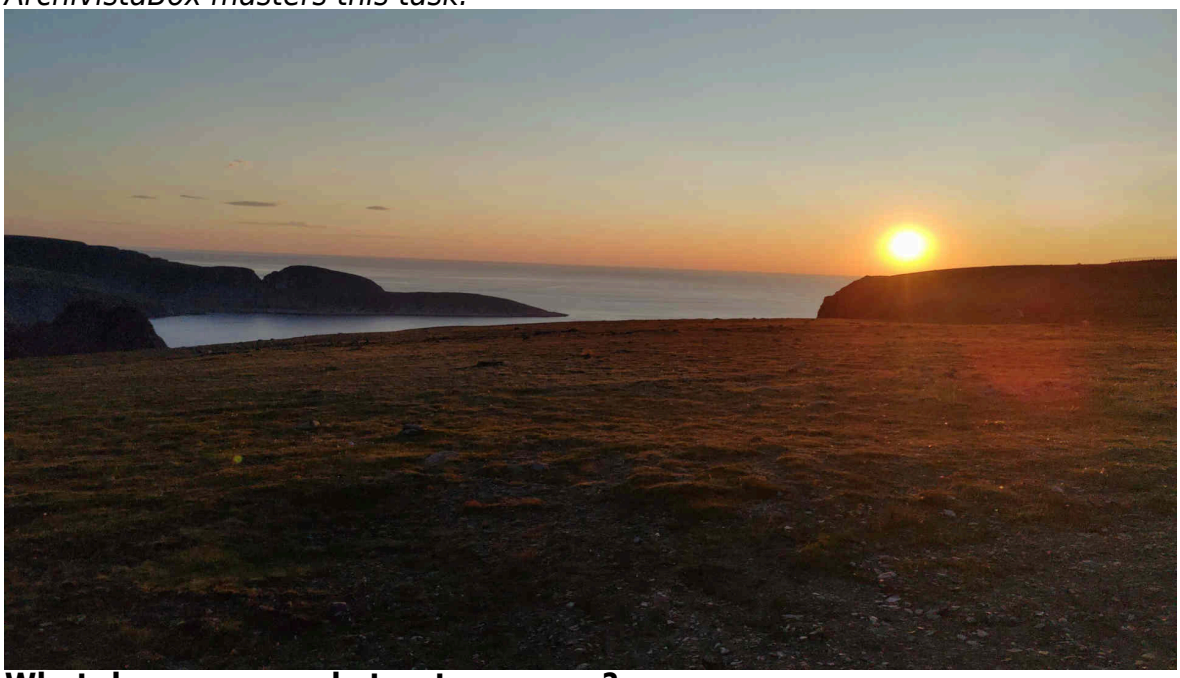
ArchivistaBox: New foundation until 2029

Egg, December 20, 2024: *Whew, this was another close call. Long planned, and yet only now ready. The new foundation for the ArchivistaBox is available from today. Little has changed for users, and not much more in terms of support. And yet the update is crucial, only thanks to a well-maintained foundation will our ArchivistaBoxes run stably, simply and securely for many years, even decades, in tough day-to-day business. This blog is not primarily about the new release, but about the complexity of digital data and how the ArchivistaBox masters this task.*



What does a new substructure mean?

With the ArchivistaBox, the solution is delivered from a single source. This means that all the required software is delivered together with the hardware, i.e. in the case of ArchivistaBox, the operating system is delivered together with the applications.

With the old release, this was either kernel 5.4 or 5.10 based on Devuan Beowulf (or Debian Buster); with the new substructure, kernel 6.10 based on Devuan Daedalus (or Debian Bookworm) is supplied. From version 2024/XII onwards, new versions will be released exclusively on the current basis; this is in line with the release in 2019. The support period for the new master release will last until the end of 2029.

This technical information is provided here for the sake of form, but is not the central topic here. During the implementation of the current foundation, the question arose as to what is central. After consulting with many customers, the decision was made that no ground-breaking innovations were needed, just continuity.

2024/XII: Stability and mass

With version 2024/XII, the scanner drivers were reimplemented and a simple remote solution was realized for all customers. The switch from PHP 7.x to 8.x, the migration of the database to MariaDB 10.x and the upgrade to KVM/Qemu 7.x were not entirely trivial. It therefore seemed more important to us to ensure that all components simply continued to run as before.

When the last major master release appeared at the end of 2019, it was all about the ArchivistaBox being able to process multimedia files. In the last five years, these possibilities have been expanded to such an extent that, looking back, it must be said that introducing something is one thing. Bringing it to life in such a way that it can cope with the rigors of everyday business is another.

In the last few days, over 6500 hours of video material were “thrown” into the archive for testing. The ArchivistaBox needed around 5 hours to process the more than 4000 files. Over 760,000 new pages were created in ArchivistaDMS. The newly recorded data volume amounted to approx. 2.2 TByte of data.
Not a single error occurred during processing. It should be noted here that the ArchivistaBox would probably not have been quite ready for this in 2019. Currently, even an ArchivistaBox Matterhorn can handle archives with several dozen TBytes, not to mention the K2 and Everest models. With the latter two ArchivistaBox models, archives of hundreds of TBytes could be managed.
And the ArchivistaBox systems are delivered on standard hardware, requiring neither a server infrastructure nor the corresponding know-how. For this to be possible, a great deal of attention to detail is required. Especially with video files, the “devil” is in the detail. To explain how complex the matter is, let’s go a little further.

It all started with the scanned image

When the Archivista solution appeared on the market in 1998, it was primarily about digitizing printed documents using scanners or adding digital images to the archive. Data that had already been captured digitally was primarily saved as Office files and ASCII text files. At that time, Office files had to be archived using virtual PDF printer drivers; in the case of text files, the problem still exists today that, in principle, nobody knows which character set was used to save the data.

Ultimately, it was and still is about obtaining images of information. ArchivistaDMS still differs from its competitors in that it does not simply store source files (e.g. a Word file), but rather the data is virtually “photographed”. This means that data can be viewed visually decades later without any problems, even if it is no longer possible to open the source file itself, for example.

XML and document containers

However, an Office document from the year 2000 does not correspond in any way to the structure of today. Whereas previously the entire content was saved in a file, nowadays individual chunks are saved. To illustrate this, let’s take the “Hello Word” example. These 12 characters (Hello World) require the following storage space in the formats (Doc-Word, Docx-Word, PDF and Tif format with 300 pixels/inch):

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

From this it can be deduced that the image copy (Tif file) is still the most efficient form (purely in terms of storage space). The PDF file requires three times more memory without solving the problem of readability over a long time horizon (e.g. embedding of fonts).

With the docx format, it should be added that this is XML data in compressed form. If the corresponding Word file is unpacked, the following files appear:

./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

And all this for a “whopping” 12 characters. Unpacked, this requires around 56,000 characters (bytes) on the hard disk. So much for the efficiency of today’s computer programs. Irrespective of this, the more data a file format contains, the more complex it becomes to keep this information alive and secure in the long term.

Video files as a complexity roller

The brief excursion into scanned data and Office formats serves as an introduction to video files. In principle, it could be assumed that video files are sequences of images enriched with some sound.

Unfortunately, the situation with video files is far more complex. In the following, we will focus primarily on the MP4 format (H264 in the narrower sense). This is also a specification (analogous to Office files) in which several chunks can be stored together. In terms of the specification alone, several video files can also be saved together. However, most players can only play the first track of video files created in this way.

In contrast, many different audio sources can be present, in varying quality, whereby it is usually a matter of recording several languages in the file. There are also subtitle files, which are integrated into the MP4 files either as bitmaps or (currently) as text files.

The Titanic steamer sinks in the middle of the file

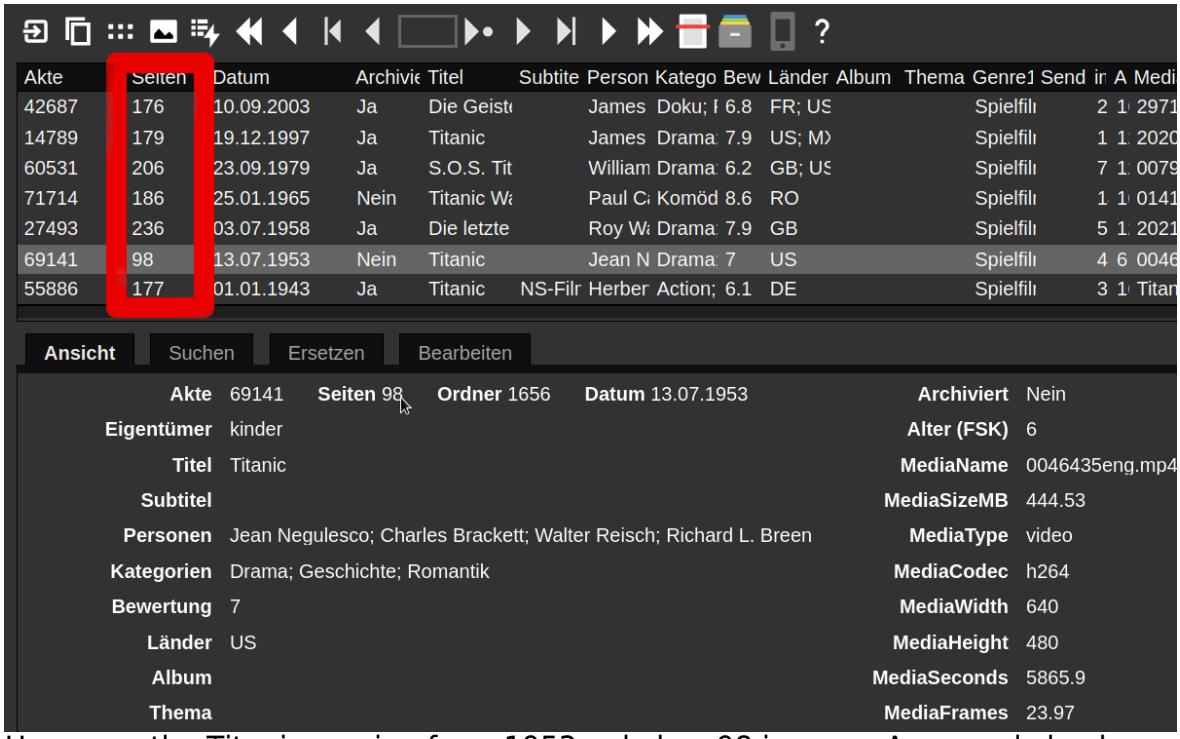
In recent years, many thousands of files have been processed for testing purposes. It was discovered that there are also video files in which, for example, the audio track is only available some seconds after the start of the video track. Good video players can cope with this, but it did not work with Firefox, for example (the audio and video tracks did not match).

In the case of subtitles, it was occasionally found that incorrect data was available. For example, in the case of the Titanic film from 1953 (which can easily be found online), it was found that subtitles from the 1998 version were included. The 1998 version lasts a whopping 3 hours and 14 minutes. In 1953, the “fight” only lasted until half-time (1 hour 37 minutes).

The main candidate for processing MP4 files (ffmpeg) sees no problem with this. The subtitle file from 1998 with time information over 3 hours can be stored in the video file from 1953 without any problems.

In practice, the 1953 version sinks “punctually” at 1 hour and 37 minutes, even if some players invite you to play it for well over 3 hours.

This was discovered by chance when the film was stored in the managing director’s private archive. This is because ArchivistaDMS creates preview images for the video files. Of course, not all images can be stored individually, as this would lead to unsustainably large volumes of data. Instead, between 150 and 300 preview images are created (except for very short video files).



However, the Titanic version from 1953 only has 98 images. A manual check revealed that the film had the wrong subtitles (those from the 1998 version, which lasts over three hours).

MP4Test: So that the Titanic no longer sails away...

The console program 'ffmpeg' is usually used to check the integrity of video files. At **superuser.com,** for example, it is recommended to check corresponding videos with the following command:

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

However, the above test does not generate any errors for the file in question, i.e. the file is found to be good. Ultimately, it is probably only if every single frame were extracted from the video that it would be possible to determine whether there are any errors. However, such a test would require a lot of computing time. Such a test would hardly be practicable for a large number of files. The little helper mp4test presented here is designed to check individual files or entire folders:

```
/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)
```

The **source code can be obtained here;** it can also be found on every AVMultimedia/ArchivistaBox version from 2024/XII. This program creates a shadow copy, whereby only 1 image per second is extracted (to speed up the process accordingly).

This means that around 15 to 20 seconds are required for one hour. The fact that the program works with several processes means that a processor with 8 cores (CPUs) can do error checking quite fast (1 minute for 10 hours). The Titanic file from 1953 can be used to show how mp4test works. Simply call up the desired file in the corresponding folder on the ArchivistaBox:

```
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
It is easy to see that a good 20 seconds were required for the test. If an error occurs, a corresponding error message is generated. The corrected version produces the following message:
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
As only one image is tested every second, the time is not 100% correct at the end. Depending on the number of images per second, minimal differences remain. Errors are always generated if the content length deviates by more than 2 seconds from the newly created shadow copy.
By specifying a directory, any number of video files can be checked in one go. With 8 cores, approx. 4,000 to 6,000 videos or approx. 10,000 hours a day can be checked.
The small auxiliary program mp4test shows well where the pitfalls of digital files lie, using video files as an example, and what effort is put into the ArchivistaBox to keep data available in the long term. With this in mind, enjoy!

Company vacations December 23, 2024 to January 3, 2025
The “official” holidays this year will last from Monday, December 23, 2025 to and including Friday, January 3, 2026. In contrast to other years, however, there is still a lot of work to be done (e.g. sending invoices) due to the extensive work on the new 2024/XII release. This year, this work will probably take place between Christmas and New Year.
Customers with maintenance contracts will of course receive support on working days (23.12, 27.12, 30.12 and 3.1). Finally, I would like to take this opportunity to thank our customers for their many years of loyalty. I look forward to being available for your requests in 2025, just as the ArchivistaBox (especially with the new master release) will of course be further developed. With this in mind, Merry Christmas and a Happy New Year.

Urs Pfister, Archivista GmbH

P.S: The pictures in this blog are from Sweden, Finland and Norway. The above picture of the managing director is from 8.8.24, after he cycled more than 4200 kilometers from Italy to Nordkapp in 19 days (on the day in question it was just under 300 kilometers). There is a documentary film about this, see ***Dall'Italia til Nordkapp.*** *All customers will receive a free online ticket for the film on request.*