AVMultimedia and ArchivistaBox as a PDF station

Egg, March 25, 2024: *A new generation of ArchivistaBox requires updated scanner drivers. We are now also making these available for AVMultimedia. The two new tools Simple-Scan and 'scantopdf' are now available to create PDF files directly on the desktop. We will also discuss the question of why JPeg compression for Fujitsu scanners has not arrived on Linux even 15 years after being financed by Archivista GmbH. And, finally, the topic of why document scanners are suddenly 4.x more expensive and what this means for the ArchivistaBox is discussed.*



Scanning, text recognition and PDFs now on the desktop

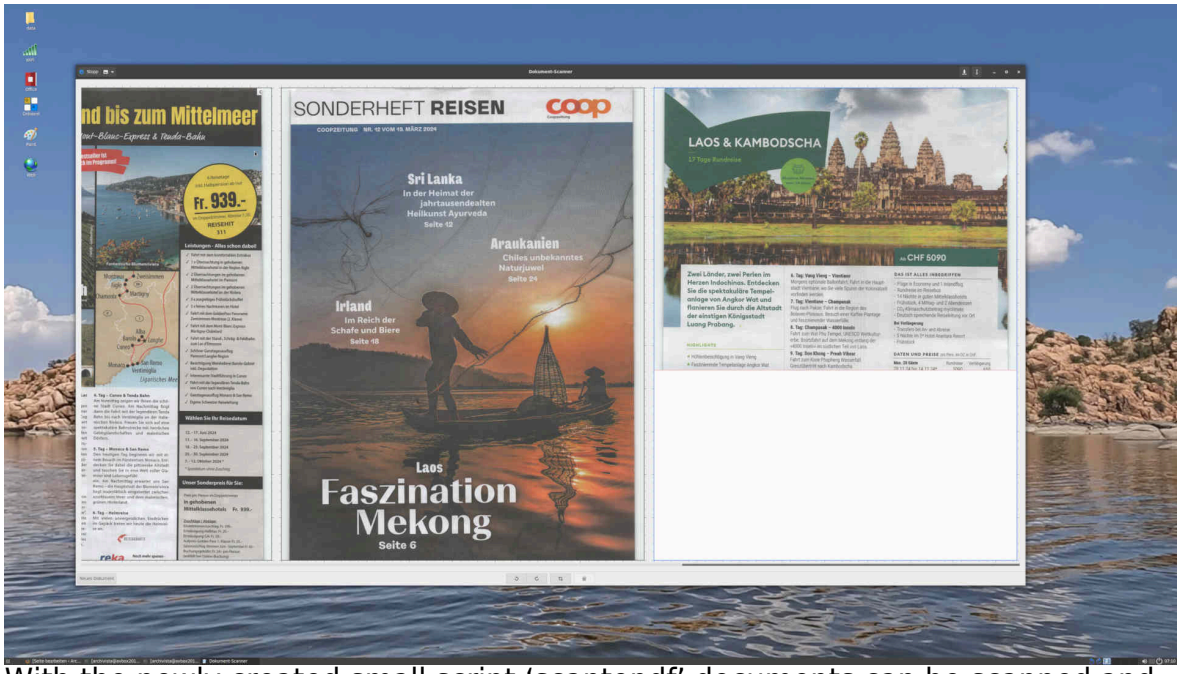
It is fair to ask why this was not already the case. Historically, it was quite difficult to work with scanners under Linux. When the ArchivistaBox saw the light of day in 2006, only a few models could be supported. For the AVision, Canon and Fujitsu devices, a mid five-figure sum had to be spent to make them work. The drivers supplied by the distributions could not be used. Worse still, the ArchivistaBox drivers were incompatible with the drivers supplied with Debian/Devuan. For this reason, the ArchivistaBox drivers were not supplied with AVMultimedia.

During these days the drivers were or had to be updated. It was found that the drivers supplied by **sane-project.org** can be used without any adjustments. Only the JPeg compression still requires a patch, more on this later.

As the standard drivers are used on the new ArchivistaBox, it is quite easy to make them available for the AVMultiedia desktop. Similarly, text recognition (Tesseract) is now so sophisticated that the standard packages can also be used here.

Simple-Scan and scantopdf as ideal little helpers

The initial situation is somewhat more difficult when it comes to selecting tools for scanning. Either these packages take up a lot of space on the desktop or testing showed that the tools were not stable. Simple-Scan is a small tool for scanning documents quite easily and creating PDF files from them. The only disadvantage is that Simple-Scan cannot create searchable PDF files by default.



With the newly created small script ‘scantopdf’ documents can be scanned and saved as PDF, with the optional flag ‘+’ also in searchable form. To capture all pages of a document scanner, for example, the program can be called on the console as follows:

```
scantopdf test+
```

The progress is logged accordingly:

```
0=>scanimage --batch=/home/archivista/data/test%04d.jpg --
mode 'Color' --format=pnm --source 'ADF Duplex' --resolution
'300' -x 210 -y 297 -l 0 -t 0 --page-height 297 --page-width
210 --buffermode=On --df-action Continue
0=>tesseract -l deu /home/archivista/data/test0001.jpg
/home/archivista/data/test0001 pdf
0=>tesseract -l deu /home/archivista/data/test0002.jpg
/home/archivista/data/test0002 pdf
0=>tesseract -l deu /home/archivista/data/test0003.jpg
/home/archivista/data/test0003 pdf
0=>tesseract -l deu /home/archivista/data/test0004.jpg
/home/archivista/data/test0004 pdf
0=>tesseract -l deu /home/archivista/data/test0005.jpg
/home/archivista/data/test0005 pdf
0=>tesseract -l deu /home/archivista/data/test0006.jpg
/home/archivista/data/test0006 pdf
0=>pdftk /home/archivista/data/test0001.pdf
/home/archivista/data/test0002.pdf
/home/archivista/data/test0003.pdf
/home/archivista/data/test0004.pdf
/home/archivista/data/test0005.pdf
/home/archivista/data/test0006.pdf output
/home/archivista/data/test.pdf
```

At the end of the process, the searchable PDF file ‘test.pdf’ is available. To obtain the options of ‘scantopdf’, the script can be called without specifying the name, the following is the output:

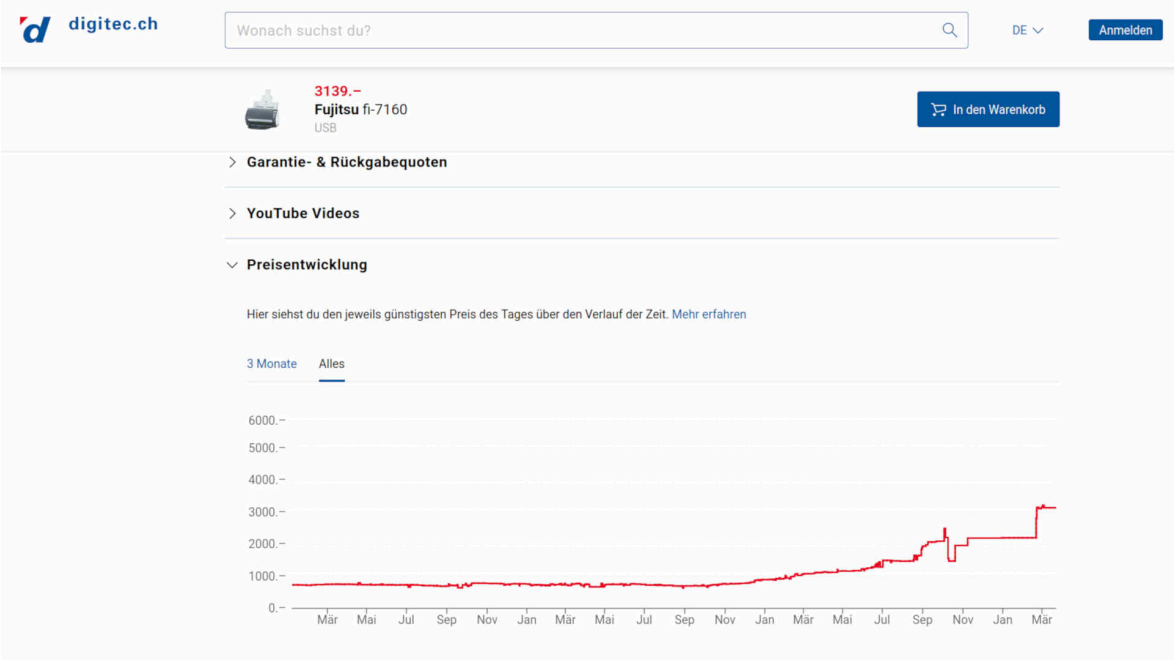
```
scantopdf file+[deu|eng] Front|Back|Duplex|Flat
Lineart|Color dpi bright
file+ means create searchable pdf or file+deu with german
language
```

Front+/Back+/Duplex+/Flatbed+[0-7] activate jpeg compression) at /usr/bin/scantopdf line 17.
If, for example, a searchable PDF file is to be created in Italian with black and white, this can be achieved as follows:
scantopdf datei+ita Duplex Lineart
It should be noted that in addition to Color and Lineart, there are also the options Gray and Halftone. Gray' corresponds to a scan in grayscale and Halftone leads to a different type of scan in black and white. Whether the 'Lineart' or 'Halftone' option leads to better scanning results cannot be determined across the board, but rather depends on the original and must be tested accordingly.



When open source (almost) doesn't come to the desktop
As a provider of a document management system (DMS), many of our customers want a document scanner to be supplied with the ArchivistaBox. A **look at our web museum shows that at the beginning of our company's history in 1998, AVision scanners were used because they were quite inexpensive at the time.** Although they also cost four-figure sums, compared to the competition, which quickly cost over 10,000 francs/euros, the AVision devices were inexpensive and solid.
The **switch to Fujitsu was made in 2008.** Primarily because these devices were able to transmit the images in JPeg format (i.e. compressed) when scanning, which meant that the pages could be transmitted much faster via UBS. To make this possible under Linux, the company Archivista GmbH invested a mid-five-figure sum to have these drivers developed under an open source license (SANE project, see sane-project.org).
However, this JPeg compression has remained deactivated in the published code to this day (for inexplicable reasons). The reason always given is that JPeg compression does not comply with the conventions of the Sane standard. Consequently, the code that was financed by our company is still not available to the general public.
The **necessary adjustments to activate JPeg compression are minimal, see here,** which should also make it clear that JPeg compression is of course activated on the ArchivistaBox or AVMultimedia.
Ricoh takes over Fujitsu and increases prices by a factor of 4.x
Ricoh has quietly "snapped up" Fujitsu's document scanner division. This affects all devices in the fi and ix series. The new owner now seems to be marketing the ix devices as consumer and the fi devices in the business sector, with price

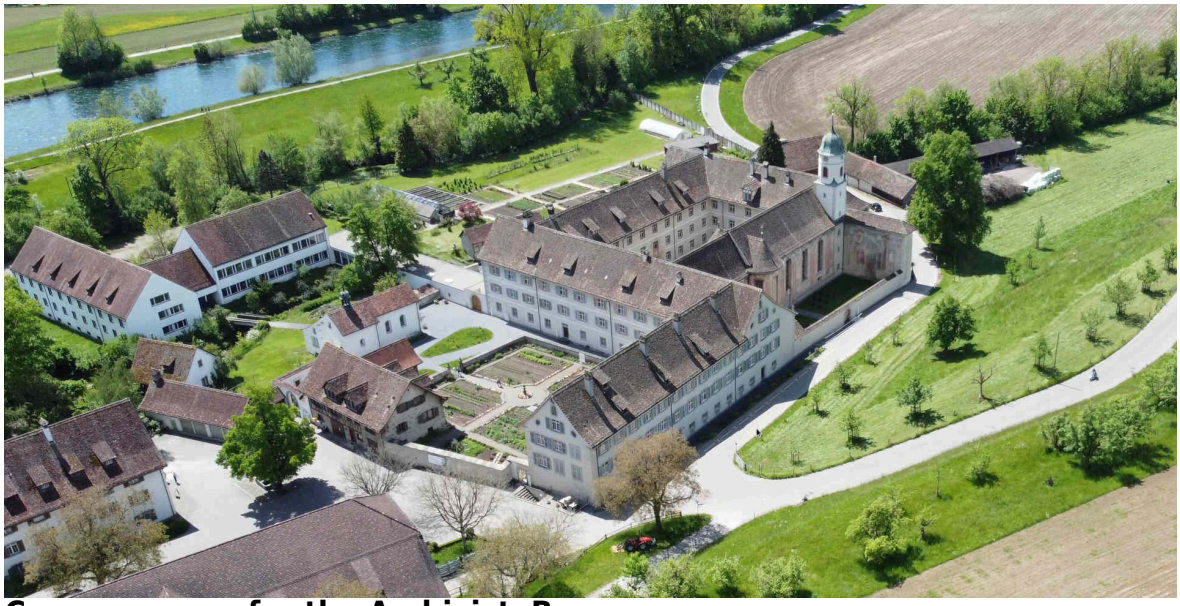
increases of up to a factor of 4.x. Whereas the fi-7160 cost around 750 francs two years ago, it currently costs more than 3100 francs/euros (according to the digitec homepage).



Opps, so the same device suddenly costs a lot more and why is the market leader Fujitsu being taken over in the document scanner sector? First of all, even if the industry repeatedly publishes forecasts that the market for scanning will continue to grow, this cannot be observed. Whereas almost 100% of documents were scanned 25 years ago, the proportion of scanned documents among ArchivistaBox customers is well below 50%, and sometimes even below 25%.
The fact is and should remain that the scanning volume is decreasing from year to year, scanners have almost completely disappeared in the private environment, any smartphone camera delivers (with a little optimization) sufficiently good results within seconds. The scanner market has been hotly contested for many years. Fujitsu has or had very good devices, but they were always a few francs above the competition.
The reasons why Ricoh has taken over 80% of the Fujitsu subsidiary (the remaining 20% will remain with Fujitsu) are not known. **However, the takeover price has been made public.** At around USD 625 million, the price seems quite modest, as Fujitsu was/is the global market leader. The takeover took place in 2022/2023 in very quiet tones, to say the least.



What’s more, the takeover is explicitly described as a rebranding (a kind of brush-up). It is true that the name Ricoh has recently appeared on the newer packaging (albeit very discreetly). However, the devices themselves only contain the model name (e.g. ScanSnap) and this is so marginal that the devices appear almost “nameless”. Or, to put it in the words of the picture below, certainly monastically modest.



Consequences for the ArchivistaBox

It should be noted that no adjustments are necessary for the ArchivistaBox or the operability of the scanners in conjunction with our scan boxes. Whether the device is from the Fujitsu era or new from Ricoh, the internal model numbers remain the same. Currently, the devices are still listed under Fujitsu in the Sane project and it can be assumed that this will remain the case for a good while yet.
The situation with the fi models is somewhat more difficult, purely in terms of the new pricing policy. The (newly) supported IX-1600 costs less than 400 francs. It can capture 40 sheets (80 pages) per minute. The fi-7160 model (previously the most widely supplied) achieves 60 sheets (120 pages) per minute. Previously the device cost around 750 francs, now it would cost around 3200 francs. This steep price difference of approx. 2500 francs is unlikely to make sense in many customer environments.
In general, it is doubtful whether Ricoh's new price strategy will be successful, as the prices of its competitors (e.g. at 60 sheets per minute) are well below CHF 1000, with the Brother ADS-4900W, for example, having a display and a network connection, while the Fujitsu or Ricoh fi-7160 only has a USB connection. Of course, this is enough to scan efficiently, but whether companies will be prepared to pay well over CHF 2000 extra for this is doubtful.
However, because Fujitsu/Ricoh prices are now so volatile, we can no longer offer them in our store. The good exception is the **Scan-Box Albis,** which now works perfectly with the IX-1600 and will be delivered accordingly. If a different device is required, we will be happy to clarify availability and prices. In addition, our customers and interested parties now have the opportunity to test the operability of a desired scanner with AVMultimedia. If the device runs there, nothing stands in the way of its use in conjunction with the ArchivistaBox.