ArchivistaBox 2015/X: text recognition, searchable PDF files and factor 2x optimisation

Egg, 7th October 2015: *Version 2015/X brings innovations that are capable of cutting the processing time for many tasks by at least 50%. The innovations allow processing in ArchivistaDMS to be spread across all the CPU cores, as required. This results in both higher processing speeds for the reading in of new documents and a significant increase in the text recognition rate (OCR). PDF documents can now be created directly in an external Windows folder, with it being possible to use the ArchivistaBox for the fully-automatic creation of searchable PDF files. Self-supporting archives can also now be created by ARM-based ArchivistaBox systems – the ISO file required for this has a size of 80 megabytes.*

Processing with as many CPU cores as required

For some years now, computers have increasingly been manufactured with multiple processors (CPUs). This can be extremely beneficial for programs and applications, but only if they have already been optimised in this regard. Until now, this has only been the case with the ArchivistaBox in respect of text recognition and on those occasions where there are many documents queueing to be processed. With the **release of Version 2015/X, the documents can now be processed in parallel by the available processors. With eight processors, for example, a 200-page document can be processed in just one-eighth of the time previously required (provided, of course, that all eight processors are actuated simultaneously).** This sounds somewhat mundane, but that most certainly isn’t the case. This is because if the total computing time is allocated to a single task, bottlenecks can then occur elsewhere.

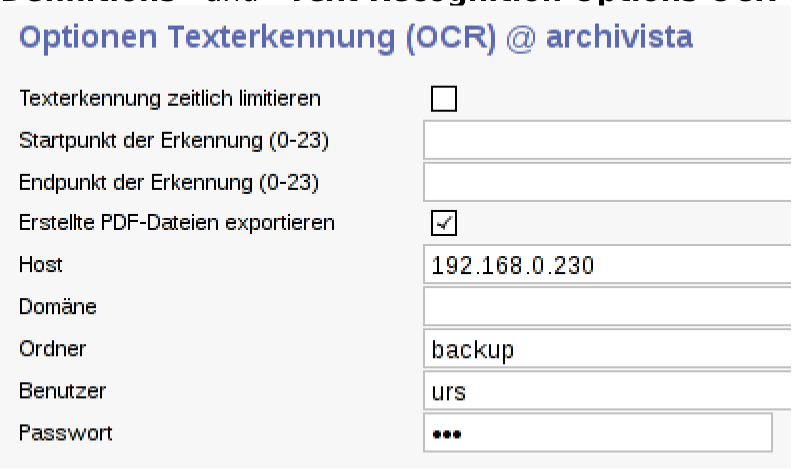
Now, the operating system monitors the ongoing applications so that no single job is allocated all the system’s resources.   In fact, the available capacity is shared out. Having said that, it is, of course, not a particularly good idea to start too many programs at exactly the same time. Again, as an example: if 1000 text recognition jobs are started simultaneously, then all the documents will be processed simultaneously, but at the expense of not being able to choose to finish specific jobs as a priority. In the worst case, if it turns out that there is too little memory (RAM) for the 1000 jobs, some of the documents will 'hang' during the processing.

The basic rule is: the number of available CPU cores must equal the number of simultaneously-running programmes or applications. This ensures that the jobs with the highest priority are executed. It is in this very regard that **Version 2015/X is strong. Depending on the usage, the CPU cores can be deployed individually per customer.** Example A: many users are accessing the archive at the same time, but the volume of documents needing to be newly recorded is fairly low. Solution A: only 1 or 2 CPU cores need to be reserved for processing purposes. Example B: a large number of scanned documents require conversion into searchable PDF files as quickly as possible. Solution B: all CPU cores are released for processing purposes. This causes archive access speed to be somewhat reduced and so the documents are actually processed more quickly (by a factor x).

To conclude, a couple of measurements from actual situations: reading the German and English versions of the ArchivistaBox handbook (PDF files with 205 / 204 pages respectively), including text recognition using Tesseract, can now be accomplished in about 7 minutes and 30 seconds by the ArchivistaBox Matterhorn. This is a performance capacity of one page per second, or some 80,000 pages per day. If documents that are already in digital format are to be processed, the **409 pages can be processed in around 50 seconds – a performance capacity of over 700,000 pages per day.**
By comparison, the "old" code needed around 19 minutes for the job with the text recognition, and around 1 minute 50 seconds to import the handbook. With the "new" code, therefore, optimisation factors of between 2.2 and 2.5 can be achieved. The use of faster CPUs would, of course, allow these performance levels to be raised even further by factors of between four and six. And using a cluster would make it possible to scale up performance levels by almost any factor desired. However, the bottom line is that **with the current code, the relevant hardware need only be capable of working half as fast in order to deliver the same level of performance.**

Creating searchable PDF documents

As has already been stated above, the current range of ArchivistaBox systems is capable of achieving very good results, particularly in respect of text recognition. ArchivistaBoxes are not only suitable for use as DMS systems, they can also be deployed as “flow heaters” for the creation of searchable PDF files. Previously, the files that were created had to be further processed using a script or the API (Application Programming Interface). Now, the searchable PDF files created can be copied directly to another network drive at any convenient time. The settings required for this can be made in **WebAdmin, under “OCR Definitions”** and **“Text Recognition Options OCR”:**



Optionen Texterkennung (OCR) @ archivista

Texterkennung zeitlich limitieren	□
Startpunkt der Erkennung (0-23)	
Endpunkt der Erkennung (0-23)	
Erstellte PDF-Dateien exportieren	☑
Host	192.168.0.230
Domäne	
Ordner	backup
Benutzer	urs
Passwort	•••

Once the option has been activated, the PDF files that are generated are saved directly into the sharing path previously specified in WebAdmin, after completion of the text recognition. If no network drive is available, the generated PDF files are saved in the TEMP folder of the Archivista shared area.

Self-supporting archives for all

Self-supporting archives can now be created with all ArchivistaBoxes (formerly Intel/AMD). Both the ARM-based and the Intel/AMD-based models now provide a “small” 80 megabyte ISO file, which allows self-supporting archives to be created. In the ARM-based model, the ISO file (archivista_cd1.iso) has to be put into the ftp/smb TEMP folder. In the Intel/AMD boxes, the file can be uploaded using the ArchivistaVM “Home” button into the folder: /var/lib/vz/template/iso. The file can be found in the “download” folder, under the name: ‘selfrun.zip’. The password for unzipping remains the same for all OS files.

Options can also be specified in WebAdmin in order to allow the archive to be space-optimised.   The image files can be compressed (a greater degree of compression is available for JPEG images) and the source and searchable PDF files can be excluded from export. This allows Archivista archives of up to 50 gigabytes in size (in the millions, in terms of pages) to be written to an ISO file and conveniently started up on an Intel/AMD computer. The self-supporting archives run exclusively in the main memory (RAM), where the size of the ISO file plus at least 600 megabytes of RAM are required for trouble-free operation.

Ready for productive deployment

The **ArchivistaBox 2015/X is ready for immediate deployment** and a current version can be requested at any time by email or telephone by customers with valid maintenance contracts. Updates can be conveniently imported using WebConfig. Thanks to the **new ARM-based ArchivistaBox range, which includes eight-core as standard, a solution now exists (for ArchivistaBox Dolder, for example) to create several tens of thousands of pages of searchable PDF files per day,** as often as is required, at a starting price of just 360 Swiss francs.