Server based OCR and document conversion with ABBYY Recognition Server 2.0

Technical overview for IT and administrators

From Pixels to Text – OCR basics

The process of transforming scanned documents to editable text, searchable PDFs or Office documents with OCR (Optical Character Recognition), takes a lot of computer resources, and it is executed in multiple steps:

- Reading the document.
- Preparing and optimising the pixel images to offer the best possible OCR results, e.g. de-skewing and removing noise from pictures.
- Analysing the documents/pages to determine the layout, e.g. headers, columns, paragraphs, lines, words and letters.
- Adaptive binarisation of colour and black/white documents before the actual OCR takes place.
- Executing the optical character recognition (OCR)
- Internal rating of the recognition results: e.g. assessment of the different hypotheses, comparing the language definitions, dictionaries and symbol examples.
- Synthesis and export to the output document formats, with or without the original formatting, e.g. basic text, office documents, export as searchable PDF files.

Technical remark:

- The OCR quality is directly proportional to the processing time.
- OCR processing always uses the maximum CPU capacity.
- All ABBYY products allow the user to set the threshold of processing speed and OCR quality, so it is possible to adjust these parameters to specific needs.

OCR processing: Desktop vs. Server usage

If OCR is only performed by certain users and at certain times it can be done on the user's own PC. This "ad hoc" scenario can be handled in a flexible and fast manner by ABBYY FineReader Professional Edition or ABBYY FineReader Corporate Edition.

In the case of higher volumes or when workgroup scanners or MFPs are used, user-focused processing is no longer suitable; without significantly increasing the effort involved in training, administration and supervision, the quality standard drops. Just as with file, print and mail servers, OCR and document conversion can be centralised and run on a server backend. Employees can use the OCR service on the network at any time and companies have the advantage of being able to convert existing paper based archives outside normal business hours.

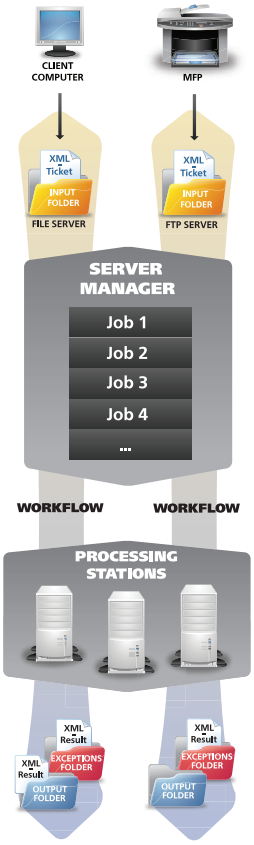
ABBYY Recognition Server was developed especially for this purpose. The system combines the highest recognition results and stability along with flexible intuitive usage, easy installation and administration. An ABBYY Recognition Server installation is also extremely scalable - without adding additional complexity. The possibility to integrate ABBYY Recognition Server in existing IT structures completes the professional requirement of ABBYY Recognition Server.

Technical remark:

- Quite often, IT departments do not allow desktop OCR applications to run on Terminal Servers; the high CPU usage can have a negative affect on the processing requirements of other users. Server based OCR perfectly compliments Terminal Server installations.

ABBYY Recognition Server Architecture

The most important terms, basic functions and the architecture of ABBYY Recognition Server are explained below:



Workflow and Job

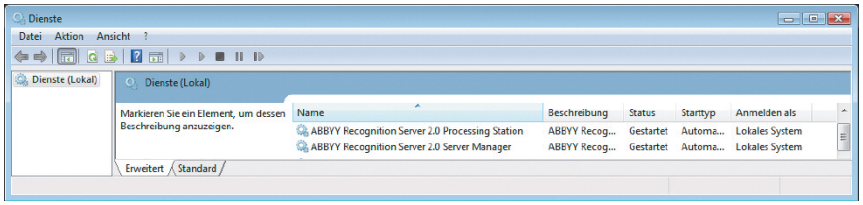
The smallest administrative unit is presented as a Recognition Server workflow. A workflow contains all processing parameters, like the source information from where documents should be processed from (e.g. FTP-, network folders or Microsoft Exchange mail accounts.) as well as the information about how they should be processed and exported. A job on the other hand is the smallest processing unit. This can be an image with one page or a PDF or image file that consists of many pages.

Server Manager and Processing Stations

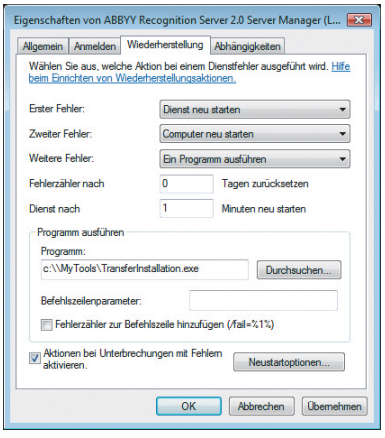
The ABBYY Recognition Server architecture is based on two main components: Server Manager and Processing Station(s):

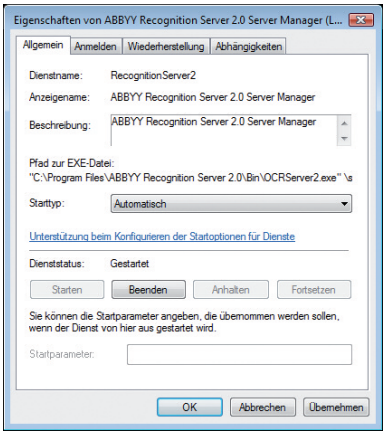
Server Manager: Administrates all processing options, monitors the dedicated folders and balances the workload by distributing the incoming files (jobs) to the available Processing Stations.

Processing Station: Executes OCR and document conversion.



Both components are implemented as Windows services. The Recognition Server services will be started automatically when the computer is booted. If a service fails unexpectedly it will automatically be restarted. It is also possible to restart the machine or launch another executable file in case of an error, so administrators have the ability to react to every incident individually. Because Windows system services can be assigned to specific domain accounts, the access to network drives can be controlled and supervised centrally. Normally the Server Manager and the Processing Stations will be installed on different machines. By default, the Named Pipes protocol is used for the communication between the components, TCP/IP can be used as an alternative if required.



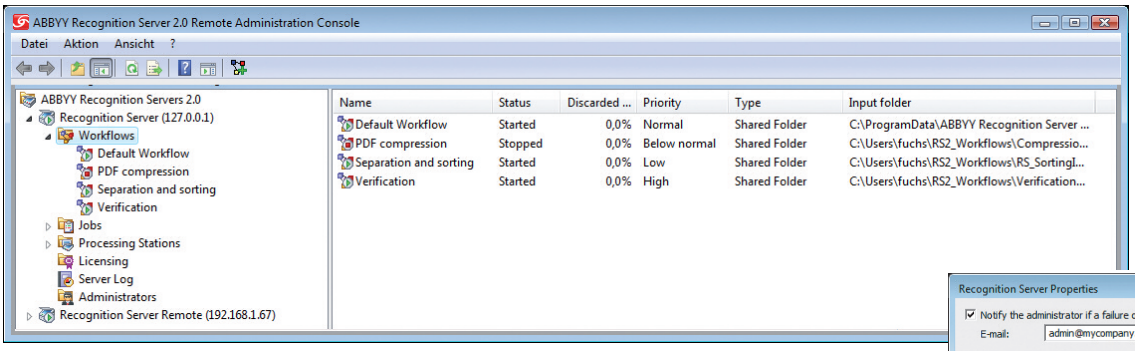


Verification / Correction Stations

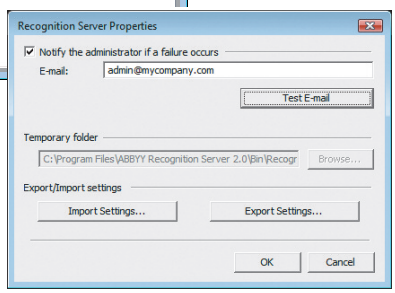
The Verification Station, a component with its own user interface, makes it possible to control and modify the recognition process to get the best results. For more detail, please refer to the section: “Complying with quality standards”.

Administration – Flexible, Remote and Central via Microsoft Management Console

Administrators can remotely manage one or more ABBYY Recognition Servers via the Microsoft Management Console (MMC). ABBYY Recognition Server settings, such as workflows, job lists, Processing Stations properties, licences, server log files, and the list of Recognition Server administrators, can be edited on a central location.



The smallest administration unit is a workflow with all its settings.

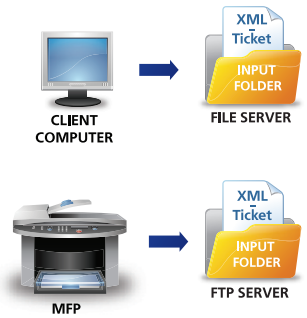


Administrators can be notified automatically via email when the Server Manager identifies an error condition.

Flexible Input Management

ABBYY Recognition Server can process many different file formats and sizes. Documents can be captured with desktop scanners, MFPs or work group scanners. ABBYY Recognition Server can process documents from network or FTP folders and can traverse any sub-folders. In addition Microsoft Exchange mailboxes can also be monitored.

ABBYY Recognition Server can handle a batch of pages and automatically merge or split documents by using blank pages or pages with barcodes and then export them as separate files.



ABBYY Recognition Server – Integration with existing software applications

Watched folders and e-mail support are not the only way ABBYY Recognition Server can be integrated. A rich set of tools allows integrators to link Recognition Server to all kinds of applications on different platforms. For example the developers can use:

- XML-Ticket Support*
- COM based API*
- Web Service API*
- Microsoft SharePoint connection*

```
- <XmlTicket>
    <InputFile Name="DemoImage1.tif" />
  - <RecognitionParams>
      <Language>German</Language>
      <TextType>Normal</TextType>
    </RecognitionParams>
  - <ExportParams>
      <ExportFormat>HTML</ExportFormat>
    </ExportParams>
  </XmlTicket>
```

XML tickets add or overwrite the default settings of a workflow.

* These functions may be available at additional charge as add-ons or part of the extended product version, depending on the country of purchase.

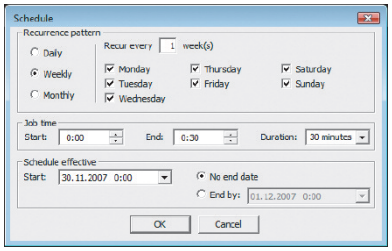
ABBYY DEFINING RECOGNITION

24/7 – Reliable processing around the clock

ABBYY Recognition Server is very stable, reliable and can handle multiple workflows at the same time.

An ABBYY Recognition Server installation normally consists of a predefined number of Processing Stations or CPU cores. Should one of the stations processing the job fail or go offline, the job will be re-routed to another available station. Naturally, the processing of the documents will be restarted and none of them will be lost during this transfer process.

To make sure that the system is not clogged by jobs that contain damaged files, any that fail to be processed will automatically be aborted and moved to an exceptions folder. Furthermore, a XML result file will be created that contains all the necessary details as to why the file could not be processed. This process also happens when a job exceeds a predefined time limit. For example, if a "normal" file takes between one and ten minutes to be processed, it makes sense to abort a job that has been processed for 20 minutes and move it to the exceptions folder.



Document routing within an ABBYY Recognition Server installation

Documents are moved from the watched folder, to a special work folder that is situated on the Server Manager computer. From there, they are distributed to the available Processing Stations. Priority and scheduler settings of the workflows as well as the total loading of the system have an impact on when each job will be processed.

After the job has been processed the original document and the files in required output formats, e.g. PDFs or Office formats, will be copied to the output folder. The actual status of a job can be checked at any time via the Administration Console.

Technical Remark:

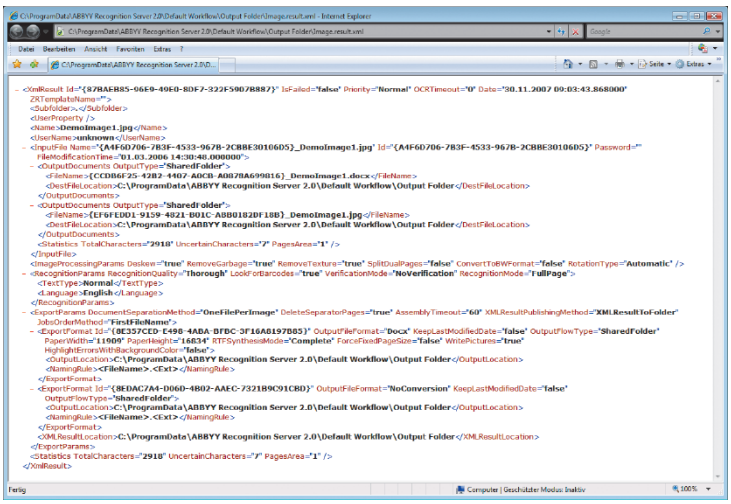
- The lowest network traffic can be achieved if you configure the watched folders directly on the computer where the Server Manager is installed.
- The Server Manager caches all the documents before distributing them to the Processing Stations, and therefore it should have enough space available on its hard drive.

ABBYY Recognition Server Log and XML–Result–File

ABBYY Recognition Server logs all jobs and errors that occur during processing in the Windows Event Log, so administrators can choose how they would like to analyse the log entries: via the Recognition Server MMC or directly via the Windows system tools.

At the job level, ABBYY Recognition Server can also create a XML-Result-File. These files contain all the processing information about the job, e.g. image pre-processing parameters, OCR Language, number of uncertain characters and so on.

When a job is aborted due to an error or manually cancelled via the Management Console, an XML-result-file will always be generated and placed in the pre-defined exceptions folder, along with the original files.



The simple structure of XML results files allows administrators to process exceptions manually or automate them through own scripts.

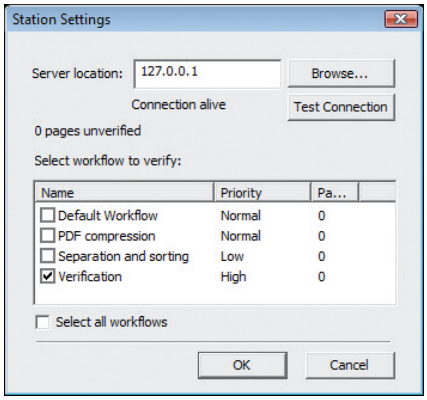
Complying with quality standards

ABBYY Recognition Server is based on the world-leading and award-winning ABBYY recognition technology, but poorly scanned documents or files with inferior quality can still contain some errors after the recognition. ABBYY Recognition Server allows defining the following quality control levels for each workflow:

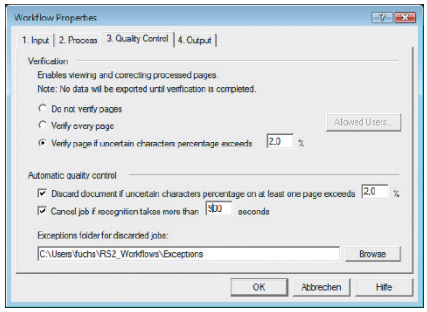
- All documents will be exported, regardless of their recognition accuracy.
- Documents that have a certain percentage of uncertain characters will be automatically rejected.
- When a defined percentage of uncertain characters is exceeded, the document will be forwarded to a Verification/Correction Station.

The Verification/Correction Station allows employees and operators to proofread the recognised text. Manual adjustments can be made to the blocks (images, text paragraphs, tables) and their position and the OCR results can also be corrected manually.

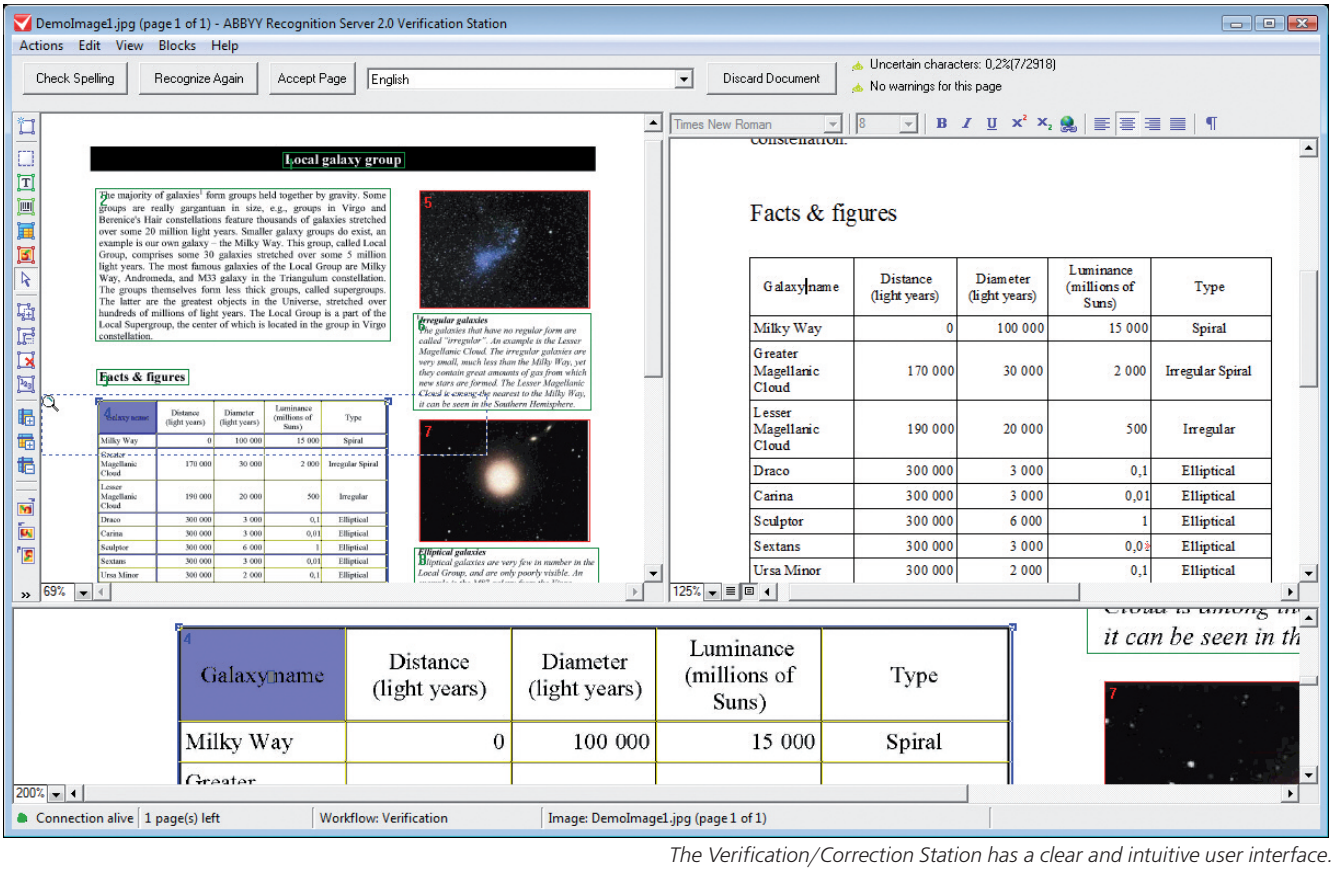
The Verification/Correction Stations can be installed on any number of workstations; the Server Manager will control the number of users that can work simultaneously according to the licensing parameters.



Users/Operators can subscribe to get pages from different workflows for correction.



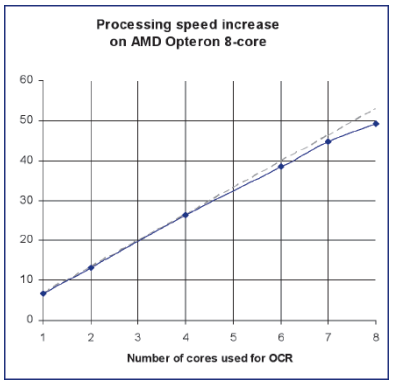
Quality and Verification settings are defined on a workflow level.



The Verification/Correction Station has a clear and intuitive user interface.

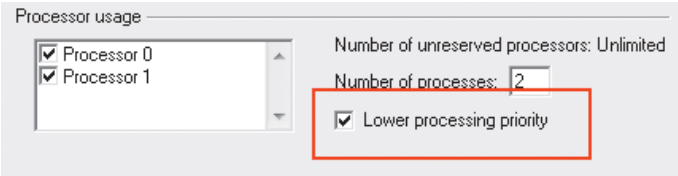
Unlimited Performance – scalable through machines, CPU cores and OCR processes

An ABBYY Recognition Server installation can be easily scaled to grow with the needs and size of an enterprise. The software is designed for the projects where the processing time is the first priority, but it is also appropriate if you have millions of documents that need to be processed continuously and reliably. Both scenarios were taken into consideration when ABBYY Recognition Server was created.



Scalability on multi–core machines

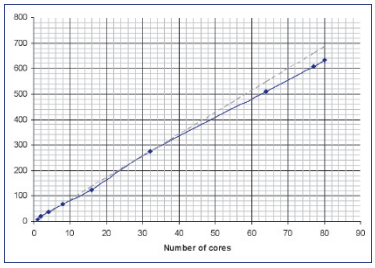
ABBYY Recognition Server Manager and Processing Station can be installed on a multi-core system and such a system is scalable in an almost linear way (see graphic). It is also possible to lower the priority on the Processing Stations to allocate more system resources to other applications in shared installations.



Scalability by using multiple machines

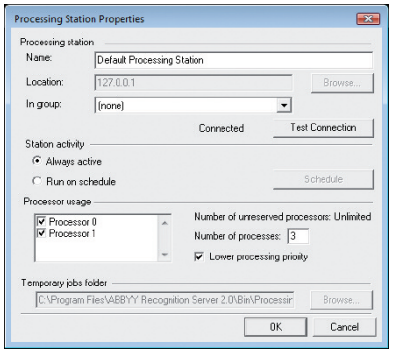
To ensure a better distribution of the work load and system stability, it is recommended that ABBYY Recognition Server is scaled on multiple machines.

The diagram shows that an ABBYY Recognition Server installation scales in an almost linear manner. The slight deviation is caused by the fact that a dual core Processing Station doesn't bring twice the performance two single core Processing Stations do.



Processing Stations and recognition processes

In addition to the licensed single and multi-core Processing Stations, the administrator can also decide how many OCR processes are started. In theory it is sufficient to start two OCR processes on a dual-core but, in practice, one sees a slightly higher performance when three OCR processes are started on a dual-core machine. When there are enough jobs in the queue, this ensures that the computer is working at full efficiency.



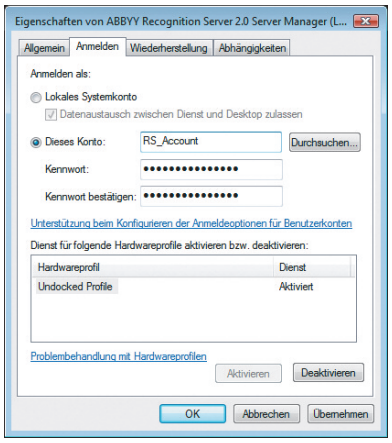
Technical remark:

- The performance in the diagrams above is based on a compilation of many thousand different documents
- The ultimate number of pages processed per minute depends on multiple factors
  - Speed of the hardware used e.g. CPUs, hard drives, network
  - Workflow settings e.g. image pre-processing parameters, OCR accuracy, export formats
  - Document structure and quality e.g. quality of the scanned document, file size, page layout, number of languages

Installation, Hardware and Operating Systems

ABBYY Recognition Server is very easy to install and a production environment can be started very quickly. To ensure that the correct access rights are given within a network, it is important that the services run under a domain account. To be able to set this up, the enterprise network should be based on a Windows domain controller. In most cases this structure is already in place and Recognition Server can be integrated without any effort.

Recognition Server can also be used in a Windows workgroup environment. As there is no domain to centralise network access, the administrators have to make sure that data exchange between network folders, the Server Manager and the Processing Stations is possible and configured correctly. For more details on installing and configuring the access rights, please refer to the Recognition Server documentation.



ABBYY Recognition Server runs on the following operating systems:

- Microsoft® Windows Server 2003/2008, Windows Vista®, Windows XP or Windows 2000
- Minimum 128 MB RAM for the Recognition Server Manager, with 100 MB RAM extra for every recognition process started on the Processing Station.
- 350 MB hard drive memory is sufficient to install the Server Manager including Processing Stations.
- Depending on the usage of ABBYY Recognition Server, the hard drive could require many gigabytes of free space as all images are temporarily saved on the Server Manager PC. To avoid the loss of data it is strongly recommended for the Server Manager to be installed on a machine with a redundant RAID system.
- The hard drive capacity for a Processing Station is a lot less as only the jobs that are currently being processed are temporarily saved there. In case of a hard drive crash on a Processing Station, there are no consequences as the Server Manager will save the jobs until they have been processed by another Processing Station.

Technical remark:

- The faster a Processing Station CPU is, the faster the document conversion and OCR will be.
- When the load is very high, the CPU load can go up to 95% for a long time.
- For stable operation of the Recognition Server installation, server hardware is preferable to a cheap desktop PC.
- Because it is possible to schedule Processing Stations, it is possible to distribute the workload in a flexible manner to all available machines

Recognition Server Licensing

The licence is always located on the Server Manager. Processing Stations can be installed without any licence. The usage/job allocation of the Processing Stationsis administered by the Server Manager.

Normally ABBYY supplies an USB licence dongle for ABBYY Recognition Server. ABBYY offers licenses which allow processing certain number of pages per month or year, or CPU-based licenses, which allow using certain number of simultaneously active dual- or quad-core Processing Stations. It is also possible to licence additional features, e.g. Verification /Correction Stations or API and integration features. The licensing policy for ABBYY Recognition Server depends on the country of purchase. Please contact your local ABBYY office or partner for detailed pricing information.