On-premise OCR service Getting Started Guide

Overview

Scan Actions in PaperCut MF offer the ability to perform Optical Character Recognition (OCR) when a document is being scanned. This allows the documents to become searchable or editable. PaperCut MF can perform OCR 'in the cloud', as part of PaperCut MF Cloud Services, or on your premises, using PaperCut MF On-premise OCR.

This document covers how to install and set up PaperCut MF On-premise OCR, which is currently in the PaperCut Percolator as Project Wollemi.

Why would I choose On-premise OCR over Cloud OCR?

Some organizations have a requirement for data to stay within their own infrastructure or even on their own premises, typically for regulatory or compliance reasons. PaperCut MF **On-premise OCR** is an alternative to PaperCut MF Cloud OCR that allows you to create text searchable documents without them leaving your site.

Be aware that this involves installing the service on selected infrastructure and keeping it updated by installing new versions.

On the other hand, **PaperCut MF Cloud OCR** takes advantage of the cloud to do the heavy lifting, removing the need for high-performance local infrastructure. This service also provides the benefit of automatically deploying service updates, so you always have the latest performance improvements and functionality.

If you're interested in learning more, there's a stack of information in the Integrated Scanning section of the PaperCut MF Manual.

Where do I install On-premise OCR?

For smaller environments, it makes sense to install On-premise OCR alongside the Application Server. In medium to larger environments, though, ensure optimum system and Application Server performance by setting up a dedicated OCR server that the Application Server can contact.

See the table below for recommendations.

Where to install On-premise OCR

Environ-ment size	Approx. jobs per day	Recommended processors*	Recommended installation location	Benefits
Small	0 – 50	2	Application Server	• Less infrastructure cost. • Great for smaller business with occasional OCR load.
Medium	50 – 200	3	Start on a well-resourced Application Server. Monitor and plan for a separate server on an as-needed basis.	• Balances resource use, system performance, and OCR processing performance.
Large	200+	4+	Separate OCR server	• Isolates resources. • Better handles high OCR load, spikes, and multiple jobs. For example, Enterprise or Education. • OCR's heavy resource requirements don't interfere with the normal operation of the Application Server.

*Recommended available processors to use (to support jobs in parallel).

Keep in mind that the more storage and processing power available, the better On-premise OCR performs—make as much available as you can. For any environment size, we recommend:

- at least 10 GB available disk space
- 512 MB available memory
- running a 64-bit edition of Microsoft Windows.

For information about:

- supported Windows versions, see System Requirements
- performance tuning of a standalone or co-located installation, see the Tune the OCR server performance section below.

Set up On-premise OCR

Install On-premise OCR

1. On the Percolator - Project Wollemi page, click **Download Installer**. The pc-ocr-server.exe file is downloaded.
   Running the installer installs the On-premise OCR service that you'll add to PaperCut MF later in the setup.

2. Run the file. The **Setup Wizard** is displayed.

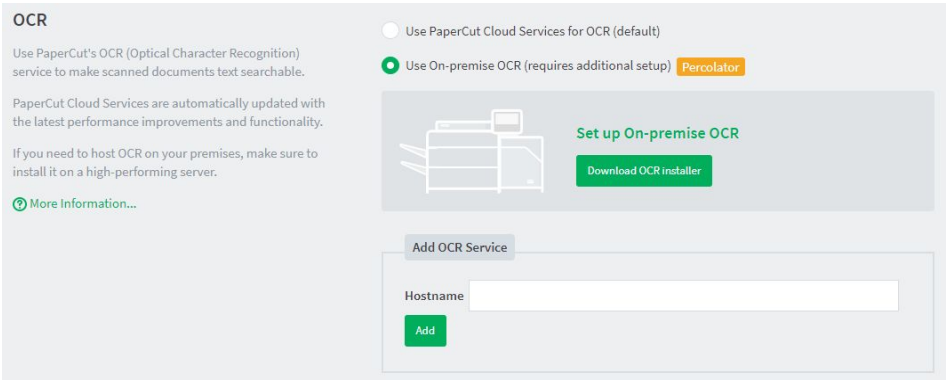
3. Follow the prompts during the install.
   - If you intend to scan documents to PDF, please ensure that the **GhostTrap** component is selected for installation.
   - If you intend to scan to DOCX, please ensure that the **Pandoc** component is selected for installation.

   When complete, On-premise OCR is installed.

   Note: The installer configures the Windows Firewall. If you are using a non-Windows Firewall, open port 9181 (inbound) to allow connections from the PaperCut MF Application Server.

Configure the host location and scan actions

1. In the PaperCut MF Admin web interface, do one of the following:
   - If you're already on the **Capture** page, refresh the page.
   - Click **Options > Capture**. The **Capture** page is displayed.



2. In the **OCR** area, in **Hostname**, type the OCR server hostname or the IP address of the server where you installed the On-premise OCR service.
   Note: We recommend that you use the IP address only if it's static. Otherwise use the hostname.
3. Click **Add**.
4. Ensure that one or more Scan Actions have been configured with OCR enabled.
5. Run a test job and check the file for success:
   - For a PDF file, check that the text in the file is text searchable.
   - For a docx file, the text should be displayed.

Tune the OCR server performance

The approach to tuning the performance of an OCR server depends on whether it is on a standalone system or co-located with other services.

By default, the OCR server processes two jobs in parallel, and they are processed with a normal CPU priority. As described below, you can change the default value of two by modifying the configuration file at `[ocr-server-path]/data/config/config.toml`.

After making changes to the config file, you'll need to restart the Windows service: `PaperCut OCR`.

Tuning for installation on a standalone system

When installing the OCR server on a standalone system, to achieve the best performance, it's a good idea to increase the number of jobs that can be processed in parallel.

The number to use depends on many factors, such as the type and size of the documents being processed and the system architecture. A reasonable starting point is to use the total number of virtual CPUs (or cores times threads on a "bare metal" system) minus two.

Put another way, if you want to process four OCR jobs in parallel and you are installing OCR on a virtual machine, give it six virtual CPUs.

To make this change:

1. In the config.toml file, remove the # at the start of the `MaxJobsInParallel` line to uncomment the option and make it active.
2. Set the `MaxJobsInParallel` line to `MaxJobsInParallel = 4`
3. Restart the Windows service: `PaperCut OCR`

Tuning for co-location with the Application Server

Note: For medium to large environments we do not recommend this approach; see the table above. OCR's heavy resource requirements can interfere with the normal operation of the Application Server.

If your system has additional available processors (beyond what the Application Server is using), you might want to consider increasing the number of jobs that are processed in parallel from the default of two.

To make this change:

1. In the config.toml file, remove the # at the start of the `MaxJobsInParallel` line to uncomment the option and make it active.
2. Set the `MaxJobsInParallel = 3`
3. Restart the Windows service: `PaperCut OCR`

Troubleshooting

After I set up the OCR server host I get an error message on the Capture page.

- Ensure that the OCR service is functioning and can be reached from the PaperCut MF Application Server.
- If you are using a firewall other than the Windows Firewall (which is configured automatically by the installer), open port 9181 (inbound) to allow connections from the PaperCut MF Application Server.
- On the Capture page, make sure there are no typos in the OCR Server hostname or IP address.
- Wait for a minute and then refresh the page to see if the error message goes away.
- Uninstall the On-premise OCR service and install it again.
- Investigate any potential network problems.

The scan worked but it's not text searchable.

- Check that the Scan Action has OCR enabled for the file type that was selected (PDF or DOCX).
- Check that the language used in the document has been enabled for OCR in the Admin web interface **Capture** tab.
- On the **Capture** tab, check the OCR section to see that the status of On-premise OCR Server is OK.

I am not satisfied with the OCR result.

Some examples of a poor result are that the whole document is partially searchable, or some parts are not searchable, or some parts don't match the original text.

- Check that the language used in the document has been enabled for OCR on the Capture tab of the admin interface.
- Try scanning again at a higher DPI. This can help when the content is particularly hard to read.
- If the document is not private in nature, or you are able to reproduce the problem with another non-private document, send the document to PaperCut support to help us improve OCR accuracy.

The test scan job failed.

- Check that the scan action has an OCR file type selected (PDF or docx).
- Try scanning again at a higher DPI. This can help when the content is particularly hard to read.
- Check that all of the pages are oriented in the same direction.
- Contact support (https://support.papercut.com/)

FAQs

Is On-premise OCR different to Cloud OCR?

As far as the output is concerned, no.

Cloud OCR accepts scan data from around the world and processes it in the region chosen by the organization. This means that data might travel outside or be processed outside the country of origin.

Cloud OCR scales according to OCR job requirements, whereas On-premise OCR requires you to manage local infrastructure and manually install On-premise OCR Server updates.

Does the On-premise OCR Service auto update?

No, not initially during the Percolator period but we're working on it.

Does the On-premise OCR Service support multiple languages?

Yes, it supports the same languages as the OCR Cloud service. You can choose to use up to 10 of the languages, however for the best balance of OCR and system performance, we recommend that you select up to five languages.

A	**F**	**L**	**S**
Afrikaans	Faroese	Lao	Sanskrit
Albanian	Filipino	Latin	Scottish Gaelic
Amharic	Finnish	Latvian	Serbian
Arabic	Flemish	Letzeburgesch	Sindhi
Armenian	Franksh	Lithuanian	Sinhala; Sinhalese
Assamese	French	Luxembourgish	Slovak
Azerbaijani	**G**	**M**	Slovenian
B	Gaelic	Macedonian	Spanish
Basque	Galician	Malay	Sundanese
Belarusian	Georgian	Malayalam	Swahili
Bengali	Greek	Maltese	Swedish
Bosnian	Gujarati	Maldivian	Syriac

Breton	**H**	Maori	**T**
Bulgarian	Haitian	Marathi	Tagalog
Burmese	Haitian Creole	Moldavian	Tajik
C	Hebrew	Moldovan	Tamil
Catalan	Hindi	Mongolian	Tatar
Cebuano	Hungarian	**N**	Telugu
Cental Khmer	**I**	Nepali	Thai
Cherokee	Icelandic	Northern Kurdish	Tibetan
Chinese - Simplified	Indonesian	Norwegian	Tigrinya
Chinese - Traditional	Inuktitut	Occitan (post 1500)	Tonga (Tonga Islands)
Corsican	Irish	Oriya	Turkish
Croatian	Italian	**P**	**U**
Czech	**J**	Panjabi	Uighur
D	Japanese	Pashto Persian	Ukrainian
Danish	Javanese	Pilipino	Urdu
Dhivehi	**K**	Polish	Uyghur
Divehi	Kannada	Portuguese	Uzbek
Dutch	Kirghiz;Kyrgyz	Punjabi	**V**
Dzongkha	Kazakh	Pushto	Valencian
E	Korean	**Q**	Vietnamese
English	Kurdish	Quechua	**W**
Esperanto		**R**	Welsh
Estonian		Romanian	Western Frisian
		Russian	**Y**
			Yiddish
			Yoruba