



# DpuScan

Janich & Klass  
Computertechnik GmbH



## DpuScan 7

PlugIn Tesseract OCR

Reference Manual

## Copyrights

© 1997 to 2024 DPU © Janich & Klass Computertechnik GmbH. All rights reserved.  
Printed in Germany. The information contained in this documentation is the property of DPU © Janich & Klass Computertechnik GmbH, Wuppertal. Neither receipt nor possession hereof confers or transfers any right to reproduce or disclose any part of the contents hereof, without the prior written consent of Janich & Klass, Wuppertal

## Trademarks

The DPU logo is a registered trademark of DPU © Janich & Klass Computertechnik GmbH. All other product names and logos are trademarks or registered trademarks of their representative companies.

## Disclaimer

The instructions and descriptions in this manual were accurate at the time of this manual's printing. However, we reserve the right to alter the description and/or the product at anytime without prior notice. As per the actual state of software technique it is not possible to develop programs that will work trouble-free under all conditions and in any configuration. DPU © Janich & Klass Computertechnik GmbH assume no liability for damages incurred directly or indirectly from errors, omissions, or discrepancies between this manual and the product.

## Actuality

It may happen that a more recent version of this manual for DpuScan is available for download from the Internet. Therefore, it is recommended that you should compare the version by means of the date printed on this page with the version on the Internet. You should please use the most up-to-date version of the manual.

# Table of Contents

<b>1 Overview</b>	<b>4</b>
1.1 Configuration in Base Profile .....	4
1.2 Configuration of the PlugIn .....	7
1.2.1 Configuration of the field search .....	8
1.2.2 Configuration of the full-text search .....	10
1.2.3 Language Selection .....	12
1.2.4 Migration von FineReader-Subprofilen .....	13
1.3 Configuration in the task profile .....	15
1.4 Configuration as a command .....	17
1.5 Displayed windows and return values .....	17
1.6 Summary .....	20

## 1 Overview

With the PlugIn Tesseract OCR a text recognition can be done in DpuScan. Optical character recognition can be carried out for sections or entire text pages. With the partial area search or field search, sections of an image are searched and the texts found there are returned in DpuScan variables. The full-text search examines the entire image and can output the results in a formatted manner. PDF, text, HTML and XML are available as export formats.

### Requirements for using the PlugIn

The PlugIn can be used in all licensed versions of DpuScan from version 6.12. No additional license is required for this PlugIn.

The PlugIn is based on the Tesseract (Apache 2.0) library. The library is installed automatically when installing DpuScan 6.12 or higher.

### Migration of FineReader profiles

The FineReader OCR used in previous versions of DpuScan is no longer offered for installation as of version 6.12. However, the PlugIn offers a migration function for the further use of such configurations (subprofiles).

### How the PlugIn works

When scanning or reading in batches of documents, the PlugIn is called up after the image data has been captured and carries out text recognition. The results are buffered in variables or special OCR files. The variables can be used by DpuScan for control or output, the OCR files are converted into the desired format together with the image.

In the interactive mode, i.e. the pause after scanning, in which the images are displayed, the PlugIn can be applied specifically to a single image.

Various configuration steps are required to use the PlugIn:

[Configuration in Base Profile](#)

[Configuration of the PlugIn](#)

[Field search](#)

[Full text search](#)

[Migration](#)

[Configuration in the Task profile](#)

[Configuration as a command](#)

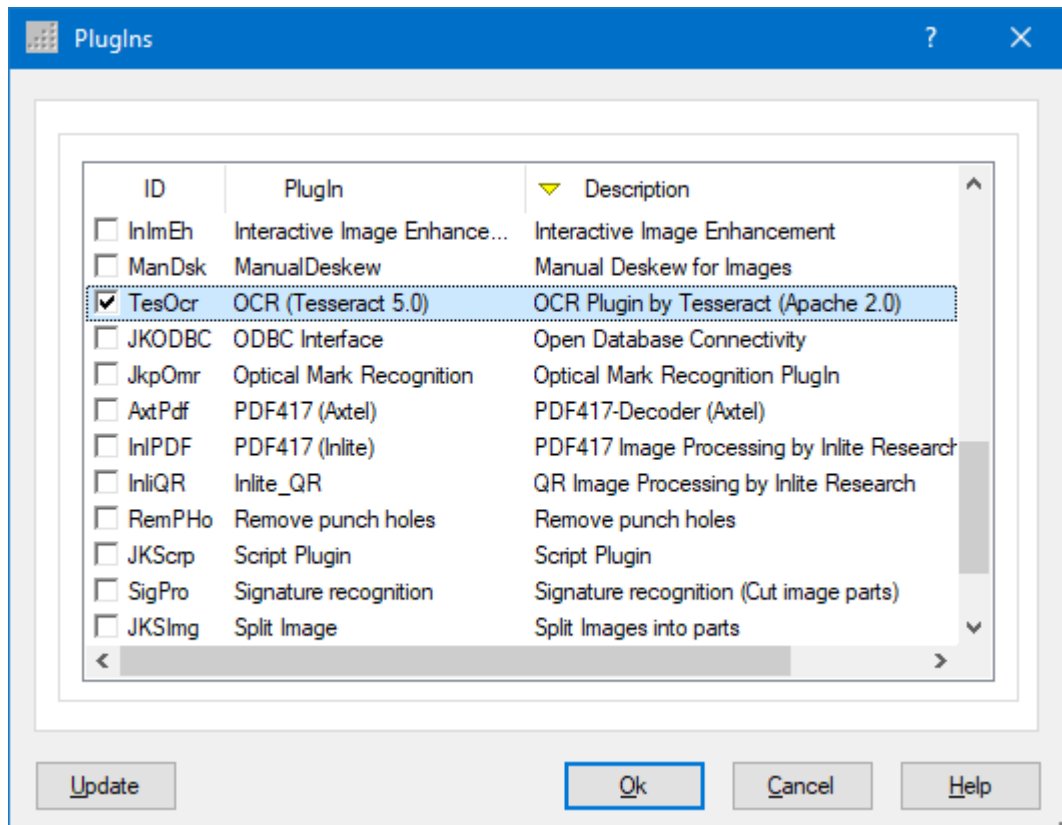
[Views and return values](#)

[Summary](#)

## 1.1 Configuration in Base Profile

The PlugIn has to be loaded and configured within the basic profile. To do this, open the **base profile configuration**, select the **Process** tab and click the **PlugIns** button.

The Add button takes you to the selection of the available PlugIns.

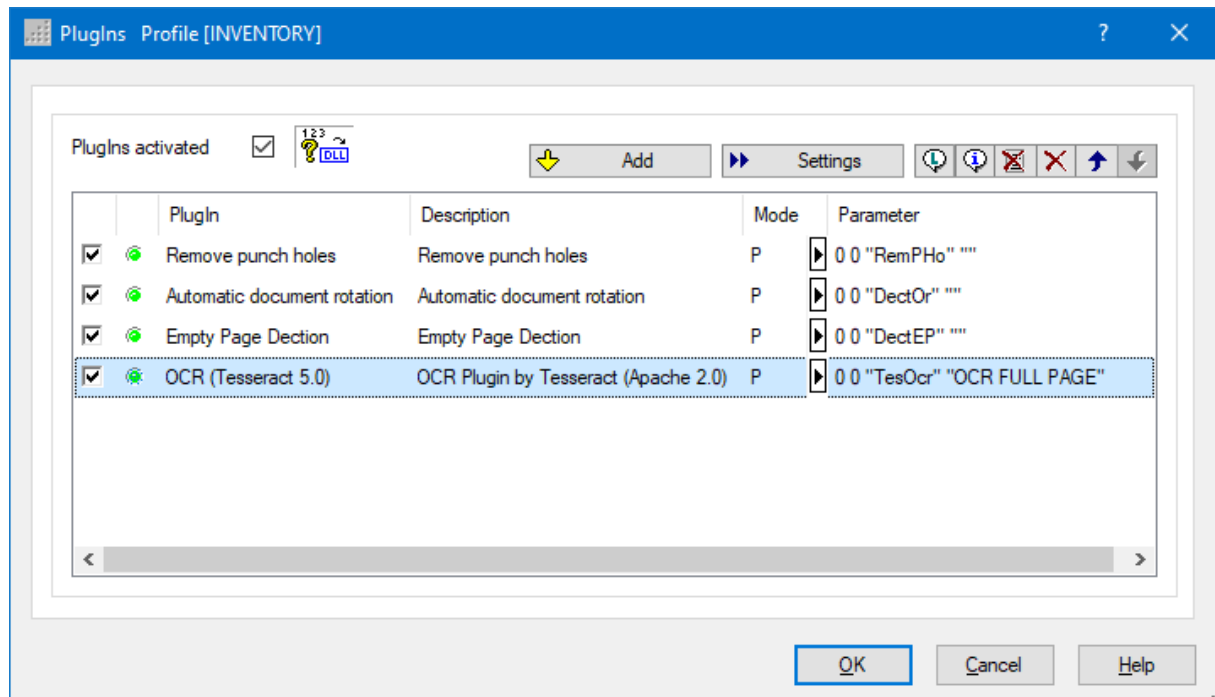


Auswahl des Plugins

Choose the ID "TesOcr" for the plugin "OCR (Tesseract 5.0)" and confirm with OK. The Plugin will be loaded for use within this base profile.

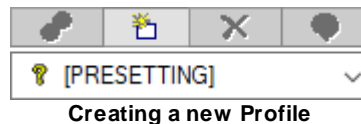
**Please make sure that the check box "Plugins activated" is checked, otherwise the Plugins will not be used. The check box can only be activated if at least one Plugin has been loaded.**

The green point in the list of loaded Plugins indicates that the Plugin is ready for use. The entry in the "Mode" column shows the string P. This means that this Plugin works in process mode.

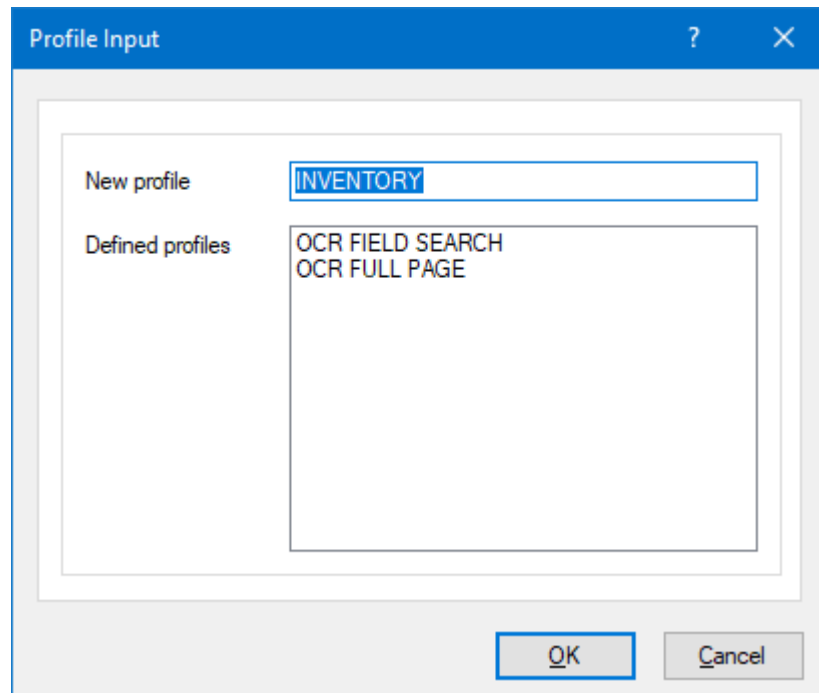


PlugIn in der Liste geladener PlugIns

Now create a configuration by double-clicking in the cell for the "Parameter" column. The dialog for calling the PlugIn as a broker event or when changing selection opens. In the middle are a number of buttons there:

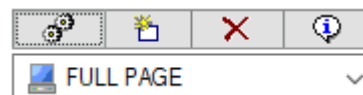


The only button available is the New button; use it to create a new sub-profile. You will be asked to enter a name for the configuration to be created:



Eingabe eines neuen Subprofils

After creating the new profile, the other buttons will become active.



Managing profiles

Click the Change button with the cogs. A separate dialog is available for the actual [configuration of the Plugin](#). After you have made the settings there, you can leave all open dialogs with OK.

## 1.2 Configuration of the Plugin

Now set the parameters for the search. The configuration dialog for the Plugin is split and shows a preview on the right. On the left are the controls for setting the search parameters.

On the the left side at the top you find drop down box for the mode

### Taskmode

Specifies how the Plugin should work:

**Field search:** Each set frame is processed individually and the result is saved in the assigned program variable ( -Code).

**Full-text search:** The entire page is recognized and the result is saved to a file.








On the right side is a **Preview Window**

Files are displayed in this window in order to test the set parameters. The image section can be moved by holding and dragging the left mouse button. The view can be enlarged or reduced with the mouse wheel. The right mouse button has no function here.

The window accepts files that are dragged onto it with the mouse. Image files are displayed, other files are ignored.

Please note that only image-only files can be displayed, files with mixed content, e.g. searchable PDF, cannot be displayed.

Above the preview window are these controls.

	<b>File open</b>	Loads a file from the hard drive and displays the image in the window. All common image formats can be loaded, such as JPG, TIF, BMP images.
	<b>Add frame</b>	Creates a frame on the loaded image. The size and position of the frame can still be changed afterwards.
	<b>Remove frame</b>	Deletes the currently active frame.
	<b>Scan</b>	Retrieve an image from the scanner.
	<b>Scanner settings</b>	Opens the scanner settings.
	<b>Test field search</b>	Performs a full recognition and saves the <a href="#">result to a file</a> .
	<b>Test full page search</b>	Performs a search for texts in all specified frames and <a href="#">displays the results in a dialog</a> .

At the bottom on the left side there is the button to start the migration.

**Migration** Opens the [dialog](#) for converting FineReader subprofiles.

Below are the usual controls.

**OK** Closes the dialog box and saves all parameters set.

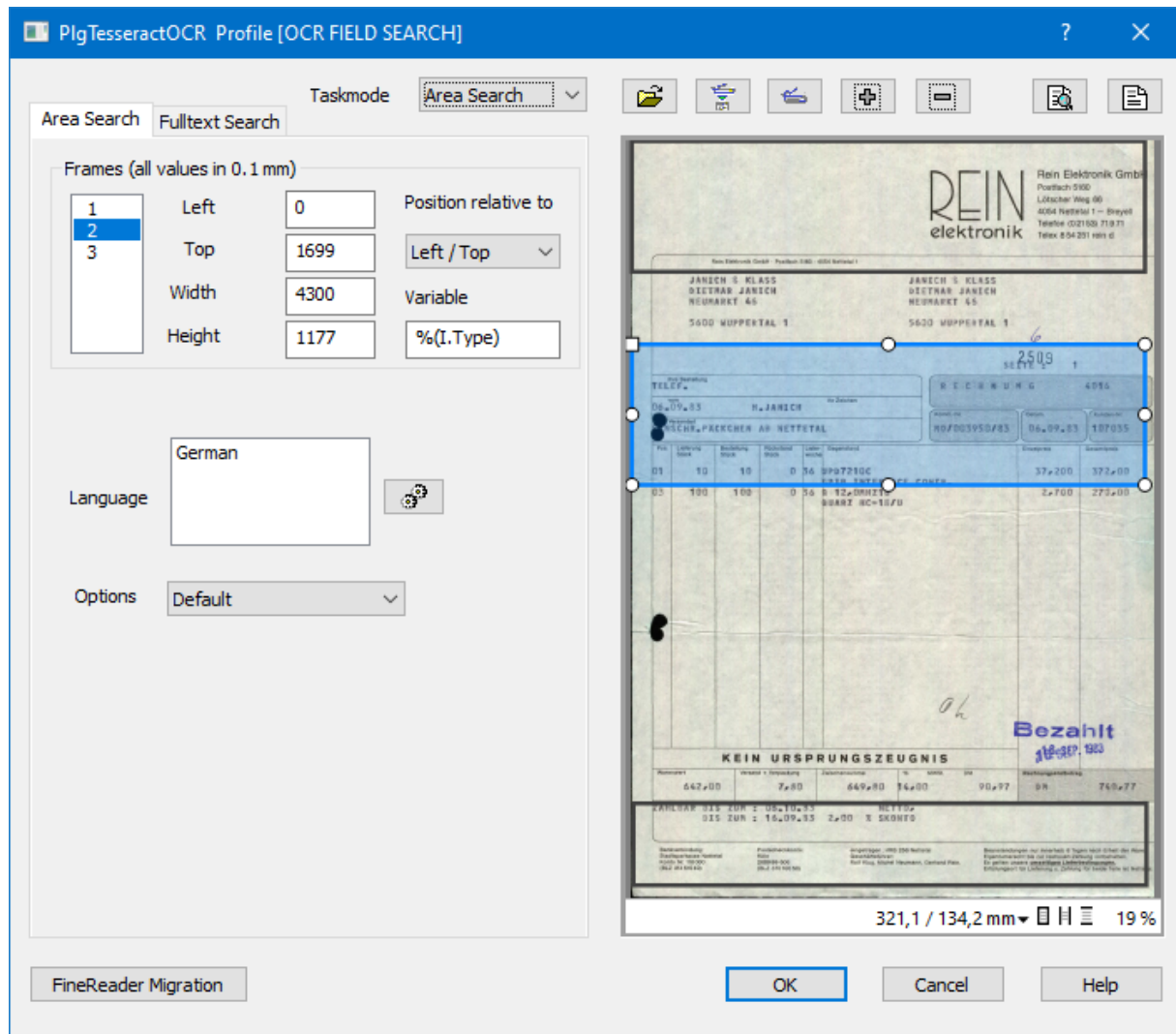
**Cancel** Discards all settings and closes the dialog box.

**Help** Opens the help for PlugIn Tesseract OCR.

### 1.2.1 Configuration of the field search

When searching for a field, text recognition can take place in one or more areas of the image. The results can be assigned to individual program variables (% codes) in order, for example, to control further workflow or filing.





Suchparameter für die Felssuche

## Search parameter

### List of the frames

Displays all frames defined in this configuration. Clicking on the corresponding entry in this list field activates the frame and displays its search parameters.

### Left / Top

An entry in this input field changes the left position of the currently active frame.

The change takes effect immediately and is displayed in the preview window.

### Width / Height

An entry in this input field changes the width of the currently active frame.

The change takes effect immediately and is displayed in the preview window.

### Position relative to

Here the recognition can be adapted to the text type:

Choose **Numbers** if only numeric values are expected in this field.

Choose **Table** if the data is arranged in a rectangular scheme.

## Language

### Language Selection

#### Options

In all other cases, **Standard** provides the best results.

Selection and order of the dictionaries to be used in the search.

Opens the dialog for [selecting the dictionaries](#) to be used.

Here the recognition can be adapted to the text type:

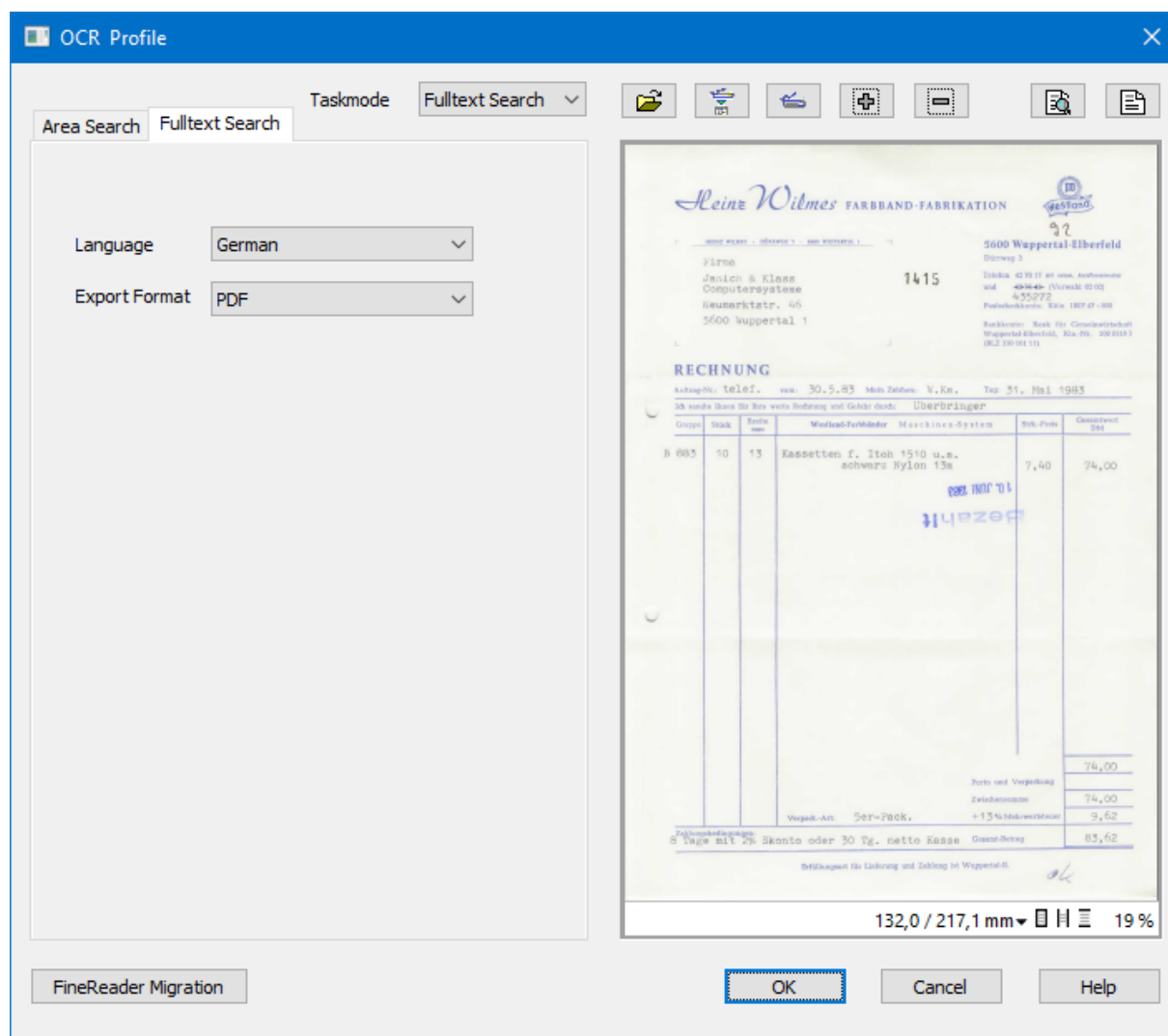
Choose **Numbers** if only numeric values are expected in this field.

Choose **Table** if the data is arranged in a rectangular scheme.

In all other cases, **Standard** provides the best results.

## 1.2.2 Configuration of the full-text search

With the full-text search, the entire image is analyzed and converted into text. Depending on the selected storage format, the image can be displayed together with the text.



Full-text Search

## Search Prormeter

### Language

Selection and order of the dictionaries to be used in the search.

## Language Selection

## Export Format

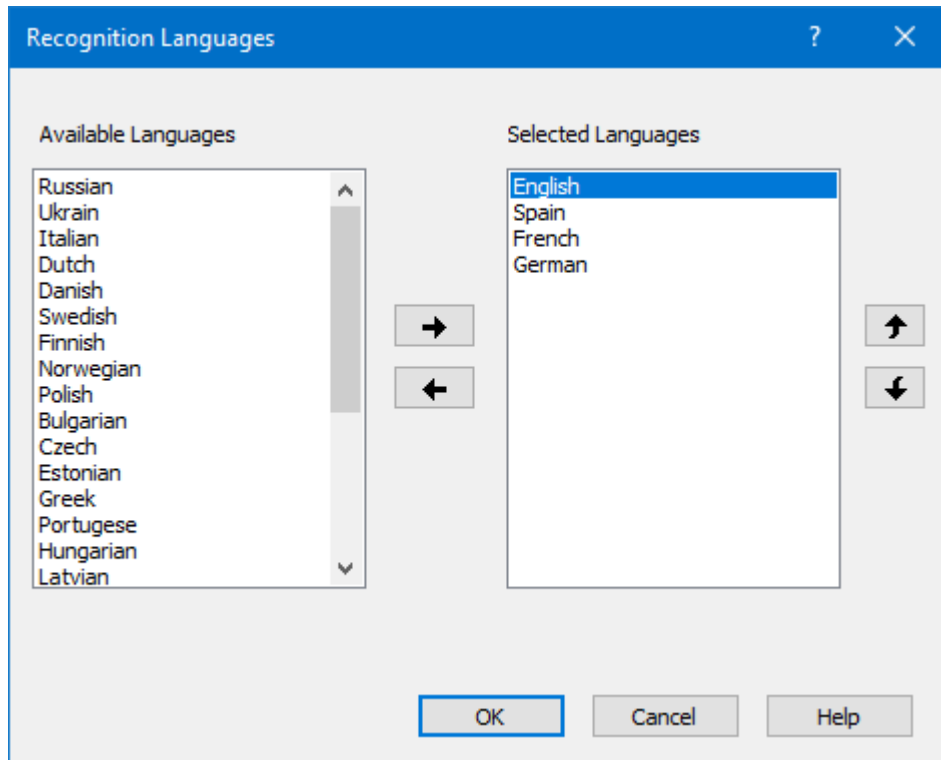
Opens the dialog for [selecting the dictionaries](#) to be used.

Specify here the format in which the text, or the text and the image, should be saved.

PDF	Creates a "Searchable PDF", i.e. the image is marked and copied.
Text	Creates a plain text file with page breaks if mul one file.
HTML	Creates an HTML file.
XML	Creates a XML file.





### 1.2.3 Language Selection

In this dialog, the **languages** can be specified and the **order** in which they should be considered. The Tesseract OCR engine uses the available dictionaries in this order to identify a word.



Language Selection

Die Steuerelemente dienen der Auswahl, bzw. der Sortierung der Sprachen:

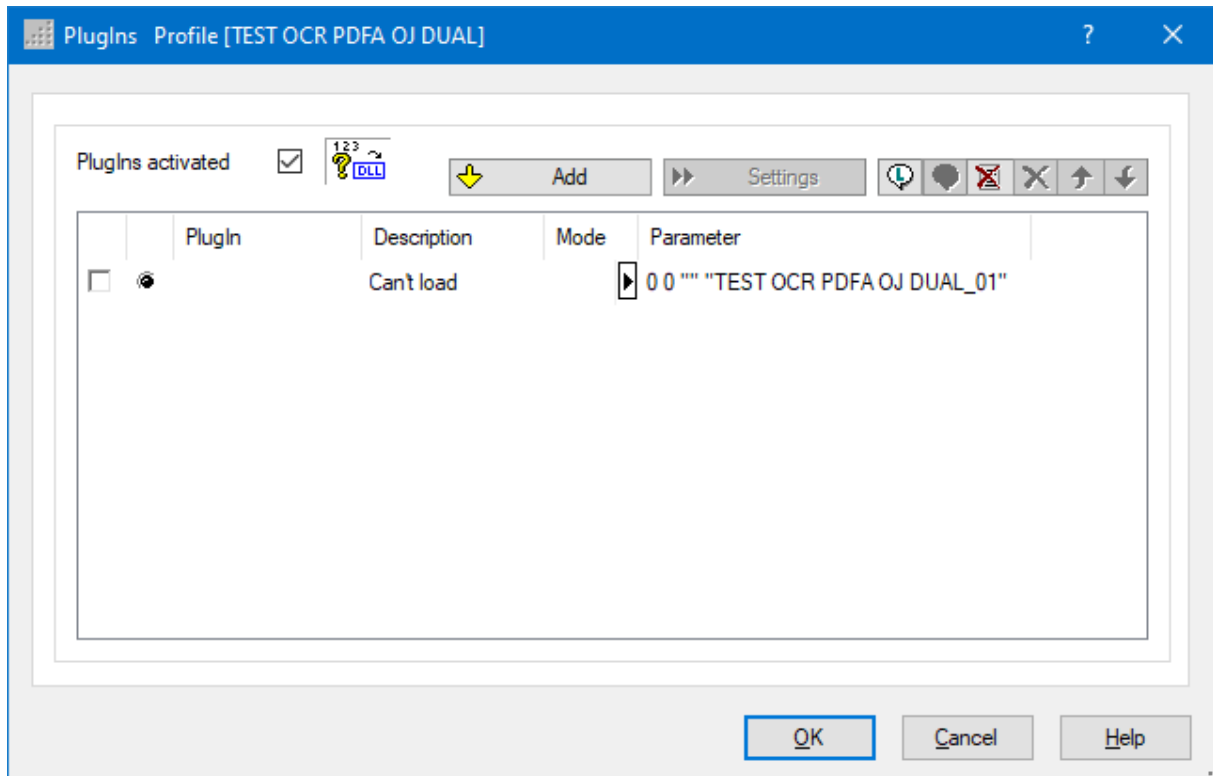
-  **to the right** Adds the dictionary selected.
-  **to the left** Removes the dictionary selected.
-  **up** Moves the dictionary up in the search order.
-  **down** Moves the dictionary down in the search order.

Below are the usual controls.

- OK** Closes the dialog box and saves all parameters set.
- Cancel** Discards all settings and closes the dialog box.
- Help** Opens the help for PlugIn Tesseract OCR.

### 1.2.4 Migration von FIneReader-Subprofilen

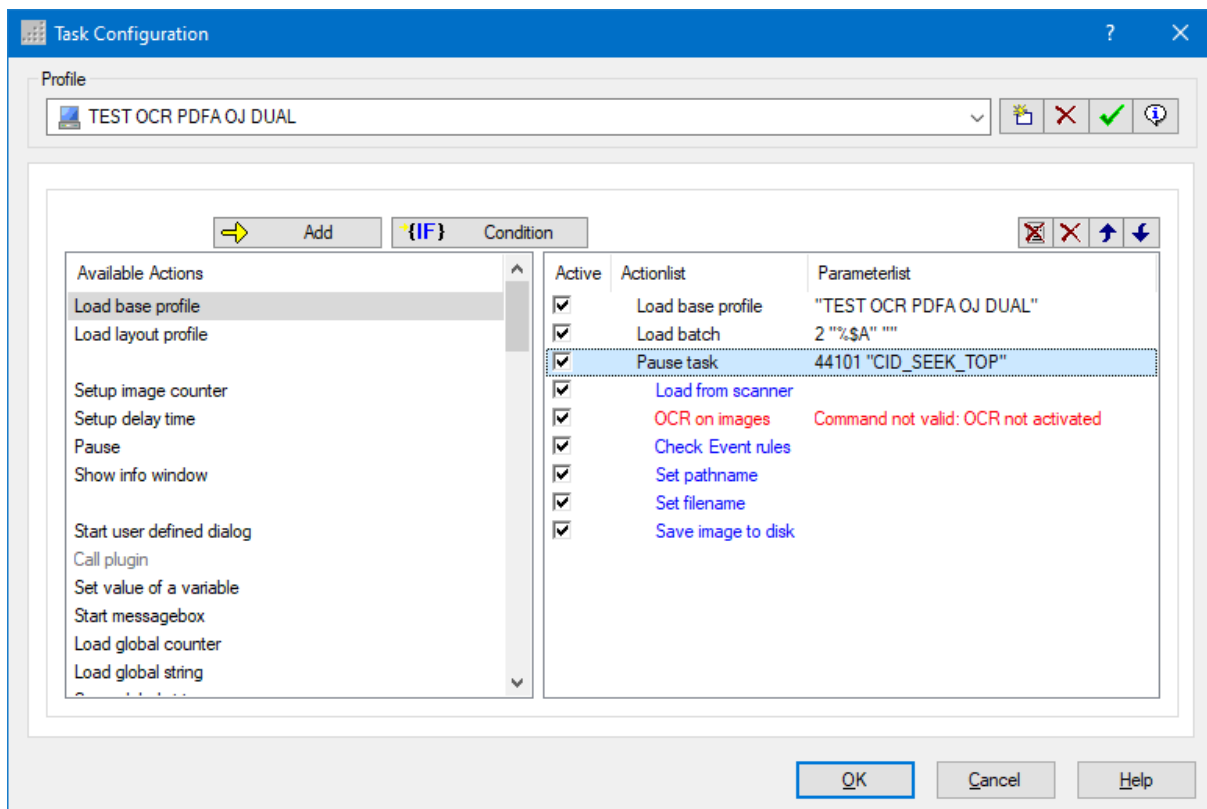
With the switch from version 6.11 to version 6.12 of the scan program, the further use of the FineReader engine has been dropped. When importing an old profile, this message appears in the base profile configuration in the list of loaded Plugins:



Unable to load the FineReader plug-in

**User here instead the *PlgTessaractOCR*.**

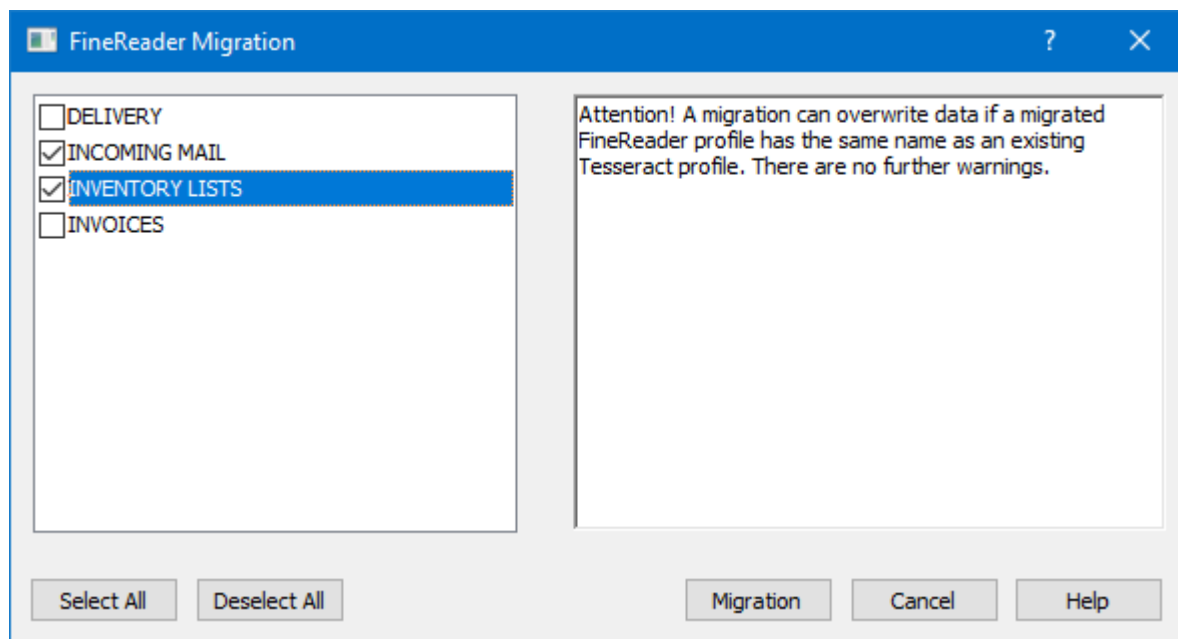
An old call to the FineReader PlugIn is displayed like this in the task profile:



Cannot use old OCR configuration

*Use here instead the task step "Call Plugin for evbery image" with the PlgTesseractOCR.*

The PlgTesseractOCR **Migration** function allows you to convert settings to Tesseract OCR format. A dialog opens with the subprofiles that could be migrated:



Migration der FineReader-Profile

During the migration, a new Tesseract profile with the same name is created and the settings are adopted if possible. The frames are always retained. The language selection also remains in most cases.

***There is no way to create a FineReader configuration from a Tesseract configuration.***

The profile selection can be set with these controls:

**Liste der FineReader-Subprofile**

List of existing FineReader profiles. This list should be checked after updating to version 6.12 or higher and each time you import an older backup.

**Warning**

As shown, if a FineReader profile of the same name is migrated, existing Tesseract profiles will be overwritten. If in doubt, you should back up all current Tesseract profiles with the DpuEnterpriseManager before the migration.

**Select All**

Selects all profiles in the list.

**Deselect All**

Cancels the selection.

**Migration**

Performs the conversion of FineReader profiles to Tesseract profiles. A new Tesseract profile with the same name is created.

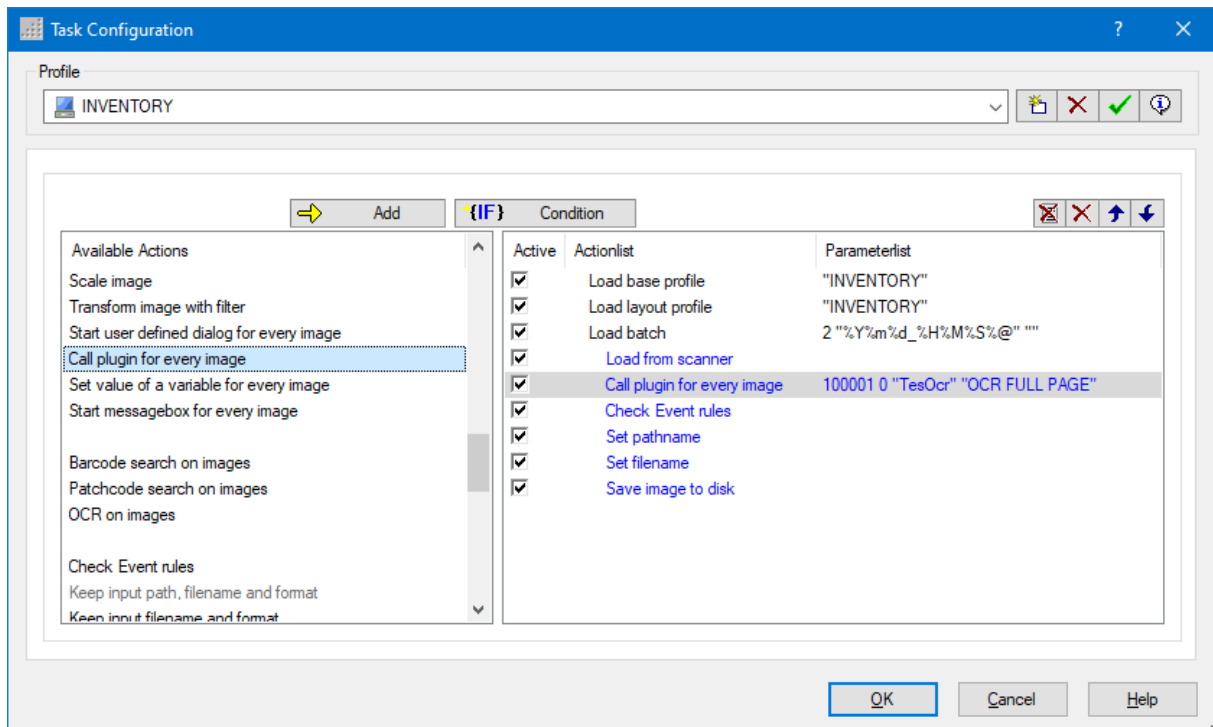
**Help**

Opens the help for PlugIn Tesseract OCR.

## 1.3 Configuration in the task profile

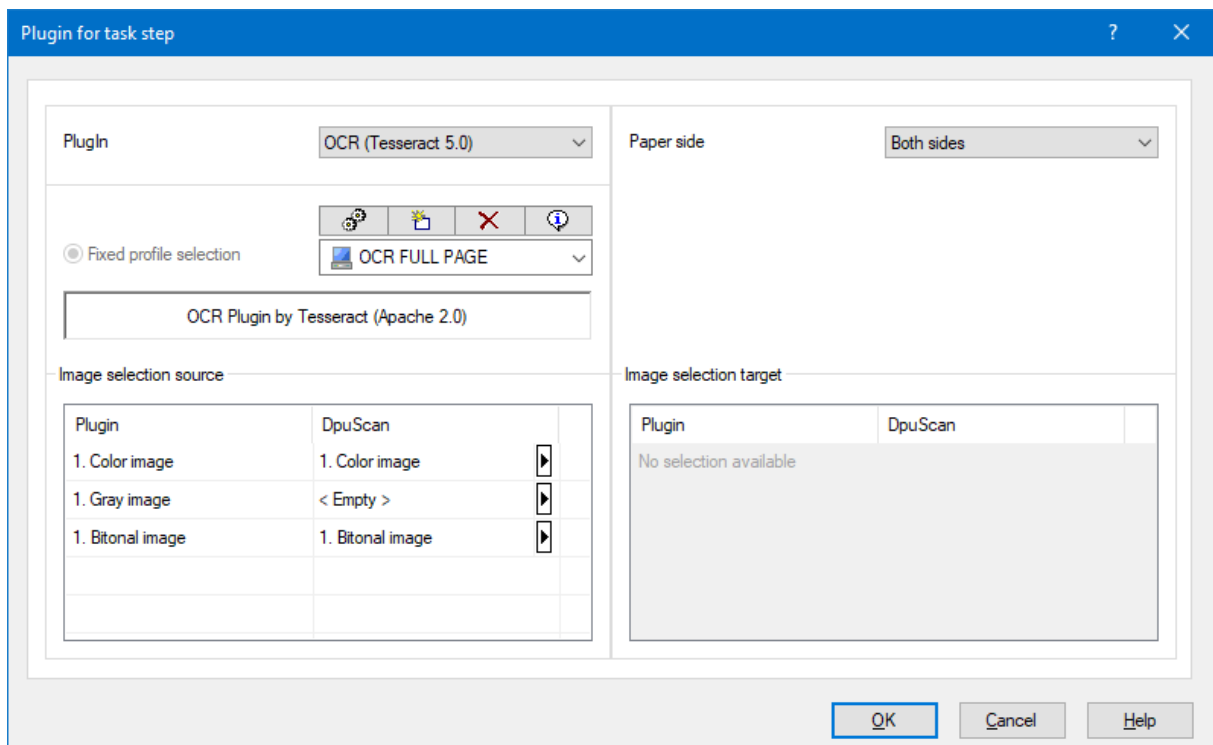
In the task profile, i.e. in the list of work instructions, the PlugIn can be inserted with the step "Call PlugIn for each image". Make sure that this step takes place after the image has been captured, here "Load from scanner". If you need the search result to control the process, it must be placed before "Execute event rules".

Remark: The instruction "OCR for every image" will work with Tesseract automatically, as soon as the subprofiles are migrated.



Call of the Plugin in task profile

Since this Plugin works with images, you must specify which images you want to work with:



Plugin für Taskschritt

The images are assigned in the dialog at the bottom left. The Plugin can only edit one color, gray or black and white image per call.



Enter which image should be transferred on the DpuScan page. Enter the image types that the scanner delivers, in the example this is only the color image.

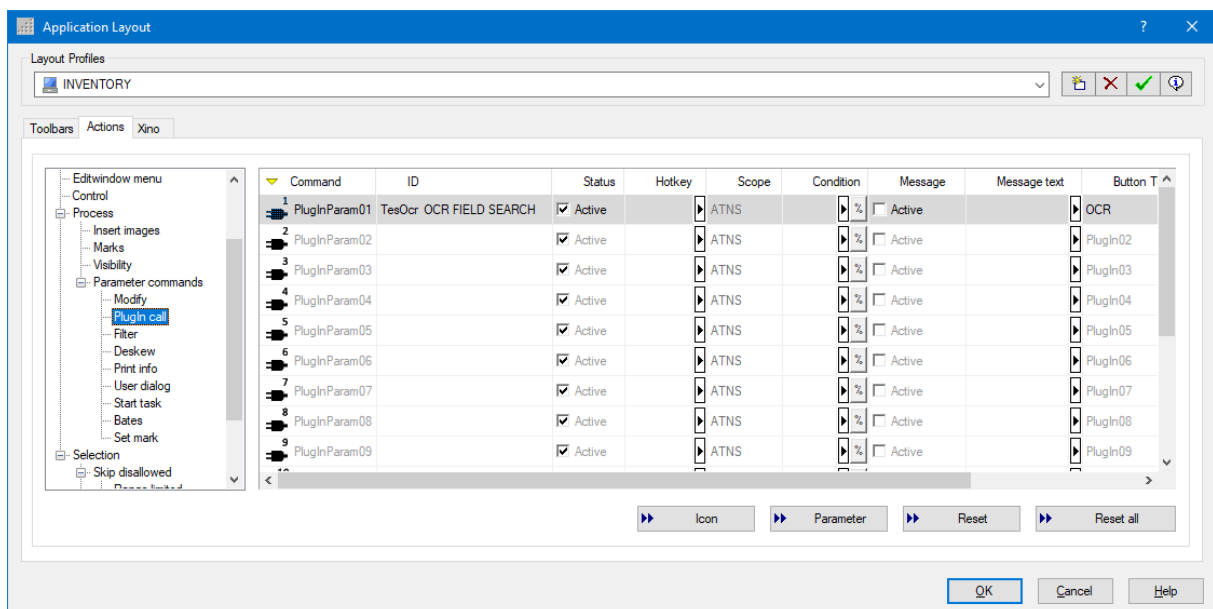
The number of the image does not indicate the position in the stack, but rather the position within an image group. In most cases the 1st image must be used.

Most recognition programs work with black and white images. If the scanner provides color and black and white images, you should pass the black and white image as a "bitonal image" for recognition.

At the top right of the dialog you can limit the search to the front page if you don't want to search on the back pages.

## 1.4 Configuration as a command

The PlugIn can also be applied specifically to a selected image. To do this, open the application layout and go to "Actions". In the tree view on the left, select the Process branch -> Parameter commands -> PlugIn call



PlugIn Aufruf als Parameterkommando

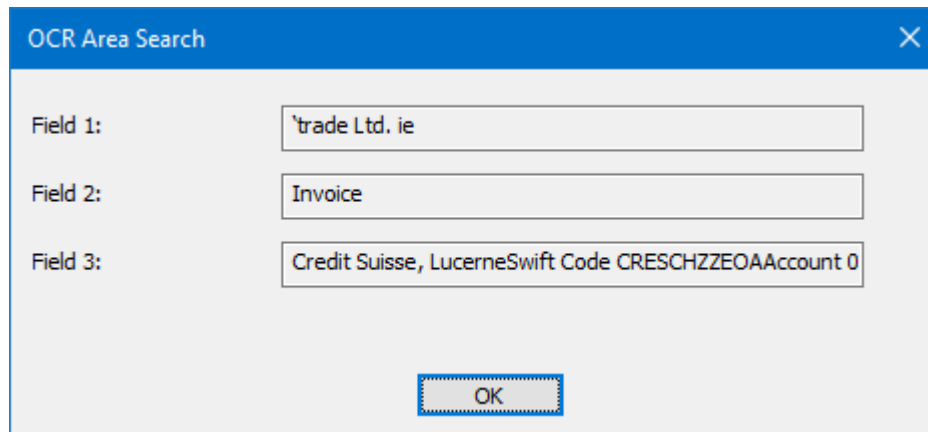
Clicking on Parameters or double-clicking on the Command column opens the [familiar dialog](#) for selecting the PlugIn, subprofile and the images to be transferred.

After specifying these values, you can assign an icon, a keyboard shortcut and various labels. Now the symbol button can be placed on the toolbar. If the keyboard shortcut is entered or this button is pressed, the PlugIn is called and the [search results](#) are updated.

If the search should be carried out as a **macro**, i.e. as part of a sequence of instructions, select user macros in the tree and insert the PlugIn call as a command. In the same macro you can then, for example, set a flag that states that the stack should be rebuilt before finalization.

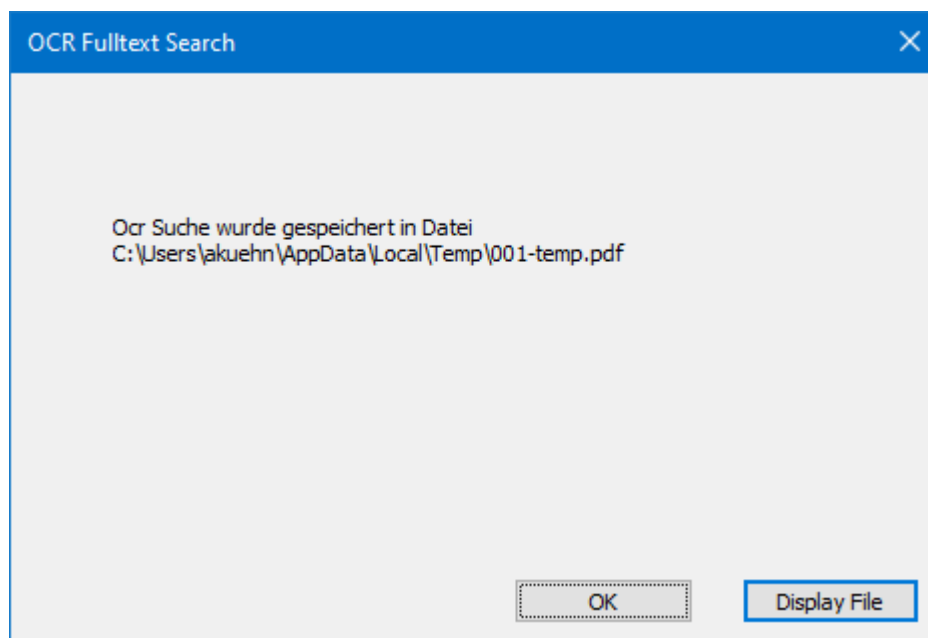
## 1.5 Displayed windows and return values

The plug-in does not display its own windows during operation. Results are only displayed as a list when the search is carried out in the configuration dialog:



Anzeige des Ergebnisses bei der Feldsuche

Full-text searches only indicate where the results were written:



Anzeige des Ergebnisses bei der Volltextsuche

### Output Variables:

<code>%(S.Res<sup>n</sup>)</code>	Text found in frame <sup>n</sup> . This variable can be replaced arbitrarily with a speaking name, e.g. <code>%(I.Address)</code>
<code>%(I.IMAGE.XMLFILE)</code>	The name of the temporary XML-file, containing the OCR result. This is intended to use for the PlugIn Classify.
<code>%(S.OCRFILE)</code>	Returns the name of the file holding the OCR-Result, if a non-

image format was selected. For example TXT, RTF or XML, if available.

## 1.6 Summary

<b>Name of the Plugin</b>	PlgTesseractOCR
<b>Description</b>	Captures text an an image.
<b>State</b>	8/14/2024
<b>DpuScan</b>	Version 6.12 and higher
<b>Plugin Files</b>	PlgTesseractOCR.dll, PlgTesseractOCR_07.ing
<b>Additional Engine</b>	Tesseract OCR, Version 5 or higher.
<b>Chargeable</b>	Yes
<b>Can be used as task step</b>	Yes
<b>Can be used as macro command</b>	Yes
<b>Can show a window</b>	No
<b>Reacts on broker events</b>	No
<b>Reacts on selection changes</b>	No
<b>Input Variables</b>	
none	
<b>Output Variables</b>	
<b>% ( S . Res <span style="color: red;">n</span> )</b>	Text found in frame <span style="color: red;">n</span> . This variable can be replaced arbitrarily with a speaking name, e.g. %(I.Address)
<b>% ( I . IMAGE . XMLFILE )</b>	The name of the temporary XML-file, containing the OCR result. This is intended to use for the Plugin Classify.
<b>% ( S . OCRFILE )</b>	Returns the name of the file holding the OCR-Result, if a non-image format was selected. For example TXT, RTF or XML, if available.

# Index

## - A -

Anzeige 17  
Aufruf als Parameterkommando 17

## - B -

Base Profile 4  
Bilder auswählen 15

## - C -

Changing sub profile 4  
Configuration 4  
Creating sub profile 4

## - D -

dictionaries 12

## - E -

English 12  
Export format 10

## - F -

field search 7  
French 12  
full-text search 7

## - G -

German 12

## - H -

HTML 10

## - K -

Konfiguration als Kommando 17  
Konfiguration im Taskprofil 15

## - L -

Language 10  
Loading the PlugIn 4

## - M -

migration 7  
more languages 12

## - O -

order of languages 12  
Overview 4

## - P -

PDF 10  
preview window 7

## - R -

Rückgabewerte 17  
Rückseiten 15

## - S -

Schritt im Makro 17  
Set the task mode 7

## - T -

Taskschritt 15

## - X -

XML 10

## - Z -

Zusammenfassung 20