Mariner Software

How can the text-retrieval (OCR) process in Paperless typically be expected to work?

Customer Service - 2024-05-07 - OCR

One of the new features in Paperless 2 is that Paperless **retrieves text from a library file** and writes it to the **OCR Text field**. Once Paperless has retrieved text (and written it to this field), Paperless will perform some additional analysis on text in the OCR Text field, to attempt to determine values to populate to specific fields automatically.

- **When a library item is first imported**. If the preference **Perform OCR on imported/ scanned files** is enabled, Paperless will retrieve text (and analyze it) automatically. Disabling this preference disables all of the processes listed in this article from happening automatically when library items are imported. By default, this preferences is enabled.
- **By using the Recognize Text command**. When a library item is selected in Paperless, there are three ways to run Recognize Text:
    - Select **Recognize Text** from the **Edit** menu.
    - Hold the **control** key down and click (or right-click) on the library item and select **Recognize Text** from the contextual menu that appears.
    - Click the **Recognize Text** button in the Paperless toolbar. This button is not displayed by default; to display it, you will need to configure the Paperless toolbar.

We provide instructions to review text Paperless has retrieved from a library item in this knowledge base article.

Process for retrieving text information from a library item.
In the first part of the process, Paperless **retrieves text** from most new library items. Paperless may retrieve text by performing Optical Character Recognition (OCR); it might also retrieve text by copying it directly from the library item. Depending on the type of file a library item was created from, Paperless will perform one of a couple different operations. Here is an outline of what Paperless does to retrieve text in each case:

Case 1: The library item was created by importing a PDF that contains searchable text.

Some PDFs contain text that can be selected. This text can also be pasted (after being copied) into a text editor (like TextEdit or Mariner Write). Frequently, this text can also be searched for within the document: if you open the .PDF file in Preview and search for a word

or phrase in the document, Preview will return results in the document that match the word or phrase you entered into the search field.

Different names are used to describe these types of PDFs. Here are some examples:

- Searchable PDF
- Text-enabled PDF
- Embedded-text (or text-embedded) PDF

If a library item was created from a PDF contains searchable text, Paperless will copy the searchable text from the PDF. Paperless will write the text information it retrieved in this manner to the OCR Text field.

Case 2: The library item was created by importing an image file.

An image file is an image file, it does not contain text that is searchable; if the file contains text, the text is image information (it is not text information).

Examples of image files include:

- An art file or digital photograph (files with extensions .JPEG, .TIF/ TIFF, GIF, and .PNG)
- Images from Web pages
- Screenshots
- A .PDF that does not contain embedded (selectable) text

If a library item was created from an image file, Paperless will perform Optical Character Recognition (OCR) on the file. Paperless will write the text information it retrieved in this manner to the OCR Text field.

Case 3: The library item was created by importing some other type of file.

Files can be imported to Paperless that are not PDF and not image files. Some examples include:

- A rich-text (.RTF) file
- A text-only (.TXT) file

If a library item was created by importing a file that is not a PDF that contains searchable text and which is not an image file, Paperless will not attempt to retrieve text from the library item. No text will appear for the library item in the OCR Text field.

Frequently Asked Questions

Why isn't it possible to select text in a PDF Paperless has performed OCR on?

A question we are asked frequently is why it is not possible to select text within a PDF once

Paperless has performed OCR on a source file.

At this time, Paperless does not add selectable text to a source file after having performed OCR. At this time, the text retrieval process only outputs text that has been retrieved to the OCR Text field.

Related Content

- Which steps can be taken to improve the accuracy of OCR results in Paperless?
- How can I view the Paperless OCR results?