Mariner Software

Which steps can be taken to improve the accuracy of OCR results in Paperless?

Jim Henson - 2018-12-17 - in OCR

Which steps can be taken to improve the accuracy of OCR results in Paperless?

These are the basic steps that the OCR process in Paperless follows:

1. The computer **acquires an image of text** (through a scanner) from an original document.
2. The computer **passes the image through an OCR engine**, where all of the following take place:
    - The OCR engine scans the image.
    - The OCR engine determines whether any part of the image appears to match characters it is 'trained' to recognize.
3. The OCR engine **produces output** of the letters it has determined likely matches for as OCR output.

In this Knowledge Base article, we provide an overview of all of the factors that affect the **fidelity** (accuracy) of text returned by the OCR process to the original document. Here are some simple ways to improve the results of OCR in Paperless:

Increase the resolution the document is scanned.

Humans often hold texts closer to their eyes (or increase the size of a font on a computer screen) in order to reveal details in whatever they are trying to look at. Similarly, increasing the resolution a document is scanned at effectively increases the font for the OCR Engine. We find that roughly **300dpi** produces the best results. There are two drawbacks to scanning at a higher resolution:

- Scans at higher resolution **tend to take more space on a hard drive** than scans at lower resolution.
- Scans at higher resolution **tend to take longer to complete** than scans at lower resolution.

Limit the number of colors that the scanner scans.

Limiting the number of colors decreases the number of visual factors that an OCR engine needs to ignore in order determine where (and what) the letters are that it should recognize. Scanning documents in either **grayscale** or black-and-white provides the OCR

engine with images that are only either light or dark--this will the only level of value (positive/ negative value) that the OCR engine to discern between, in-order to recognize text from noise. There are two different settings that **Image Capture** provides that we find produce better OCR results:

- **Text** - processes the scan job as two colors: black and white.
- **Black and White** - processes the scan job in grayscale.

ScanSnap Manager provides two settings that we find produce better OCR results:

- **B&W** - scans a document in two colors: black and white.
- **Gray** - scans a document in grayscale.

Related Pages

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