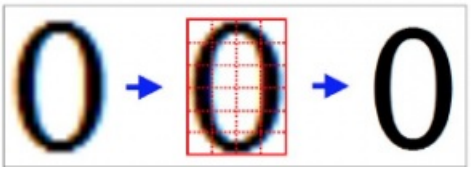
Making accessible scans

Last Modified on 06/27/2024 3:04 pm EDT

The Canon copiers at Bryn Mawr College automatically apply English-language Optical Character Recognition (OCR) to new scans and other campus tools can successfully translate OCR documents in other languages. However, OCR only works if the scans are of decent quality. This article explains how to create scans that can be successfully converted by OCR software.

What is Optical Character Recognition (OCR)?

A scan is simply a photograph of a page. The textual elements visible in that photograph are not editable, searchable text — they are simply patterns of light and dark pixels. In order for readers to read a scan using text-to-speech or Braille software, highlighting and annotation tools, and other assistive technologies, these patterns need to be converted to actual text — that is, to a string of characters that can be highlighted and searched — through a process called optical character recognition (OCR). The OCR software looks at the patterns of lights and darks and uses algorithms to determine which patterns are most likely to be characters and which characters they are most likely to be.



The output of an OCR conversion is only as good as the input. If the OCR process can't correctly identify and interpret characters, the text it generates will be nonsense and the PDF will not be accessible.

Making scans that OCR correctly

- **Start with a clean original:** Highlighting, underlining, and page damage are primary culprits in preventing the OCR process from properly recognizing text.
- **Avoid marginalia:** Marginalia can also confuse OCR software, producing extraneous characters and interfering with its ability to correctly predict and interpret neighboring words. For best results, erase marginalia or find a clean copy.
- **Keep the page straight:** Scan with all pages oriented in the same direction and as close to horizontal or vertical as possible. Most OCR tools process can correct for slight skewing, but text on pages that are highly tilted will not be interpreted correctly.
- **Don't block the text:** Avoid cutting off text or blocking it with your hands, bookmarks, etc. OCR software uses natural language processing to analyze text. Not only is the missing text not recognized or read, but its absence prevents OCR software from recognizing the missing text, but also interferes with the OCR software's ability to accurately infer and interpret the surrounding text.
- **Scan only one page at a time:**

- Most OCR software can recognize that documents scanned “two-up” — that is, with two facing pages in a book or journal scanned at the same time — have two columns of text. However, two-up scanning often creates shadows and distortions that can prevent parts of the text from being correctly interpreted.
- If each page of your original has multiple columns of text, you must scan one page at a time.

Complete the process in Adobe Acrobat Pro

If you scan the document with the Canon multifunction printer/scanners it will OCR the text automatically. However, it will not ensure that the text is read out in the right order. (For example, if your scan contains two columns, assistive technology might read the text straight across both columns.)

In order to ensure your text is read correctly, you need to open the document in Adobe Acrobat Pro (available on all college-owned computers) and follow this two-step process:

1. Run the Make Accessible wizard.
2. Check the reading order manually.

For complete instructions on how to run the Make Accessible wizard and check the reading order in Adobe Acrobat Pro, please read Adobe Acrobat: Make PDFs Accessible.