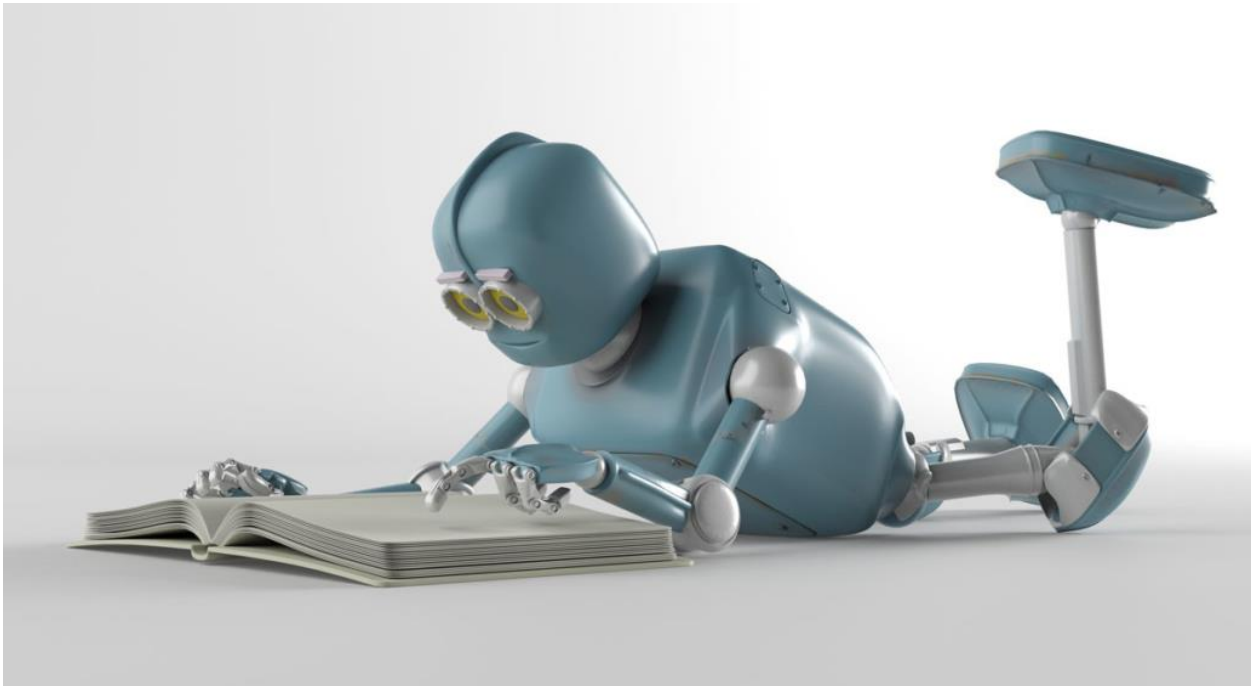
What Is OCR?

What Is Optical Character Recognition (OCR)?

How does a collection of documents transform from folders of scanned images to a complete and searchable digital library? OCR is the key to bridging high quality imaging and rich textual data.

Optical character recognition (OCR) technology allows for the conversion of scanned documents and images into a layer of text while maintaining the original image. The extracted text layer is embedded in the original image, adding searchability and research functionality.



When digitizing collections for reference and research purposes, searchable text should be considered a priority. It allows for full-text search results, which enhances the final collection when paired with indexing and metadata (information about the text, such as author, date, etc.).

Because OCR technology “reads” a digital image for text, OCR accuracy is dependent on quality scans. Poorly scanned images can cause problems when rendering search results since incorrectly processed characters or formatting will negatively affect the accuracy of the text.

How Do Digital Archivists Use Optical Character Recognition?

Starting with a batch of properly scanned images, an archivist opens the newly scanned files in the OCR software and prepares each page to be processed. This includes flagging any irregular layout scenarios like text boxes, page numbers, or graphics and

pictures that OCR can't meaningfully process. The software then "reads" all the text in the image, leaving a text copy on one side and the original image on the other.

Once the document reaches this step, the digital archivist will confirm that all formatting and structural notes were translated correctly into the text copy. During the reading process, the OCR software identifies anything not found in its dictionary, such as typos or special characters. Most OCR software includes an array of settings that allow for granular adjustments depending on the needs of a collection. For quality assurance, the archivist checks each flagged item and ensures that the text accurately matches the original.

The digital archivist can then save the processed OCR file as a variety of file types, depending on the individual collection's needs. Plain text files, HTML files, and PDFs all serve different purposes in the grand scheme of a digitization project.

I

MR. SHERLOCK HOLMES

MR. SHERLOCK HOLMES, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band, nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry—dignified, solid, and reassuring.

"Well, Watson, what do you make of it?"

Holmes was sitting with his back to me, and I had given him no sign of my occupation.

"How did you know what I was doing? I believe you have eyes in the back of your head."

"I have, at least, a well-polished silver-plated coffee-pot in front of me," said he. "But, tell me, Watson, what do you make of our visitor's stick? Since we have been so unfortunate as to miss him and have no notion of his errand, this accidental

1

I¶

MR.·SHERLOCK·HOLMES¶

MR.· SHERLOCK· HOLMES,· who· was usually·very·late·in·the·mornings,·save·upon·those·not·infrequent·occasions·when·he was·up·all·night,·was·seated·at·the·breakfast I·stood·upon·the·hearth-rug·and·picked·up·the·stick which·our·visitor·had·left·behind·him·the·night·before.·It·was·a·fine,·thick·piece·of·wood,·bulbous-headed,·of·the·sort·which·is·known·as·a·"Penang lawyer."·Just·under·the·head·was·a·broad·silver band,·nearly·an·inch·across.·"To·James·Mortimer, M.R.C.S.,·from·his·friends·of·the·C.C.H.,"·was·engraved·upon·it,·with·the·date·"1884."·It·was·just such·a·stick·as·the·old-fashioned·family·practitioner used·to·carry—dignified,·solid,·and·reassuring.¶

"Well,·Watson,·what·do·you·make·of·it?"↵

Holmes·was·sitting·with·his·back·to·me,·and·I¶ had·given·him·no·sign·of·my·occupation.¶

"How·did·you·know·what·I·was·doing?·I·believe·you·have·eyes·in·the·back·of·your·head."¶

"I·have,·at·least,·a·well-polished·silver-plated coffee-pot·in·front·of·me,"·said·he.·"But,·tell·me, Watson,·what·do·you·make·of·our·visitor's·stick? Since·we·have·been·so·unfortunate·as·to·miss·him and·have·no·notion·of·his·errand,·this·accidental¶

How to Determine If a Collection Needs OCR

How does a collector know if their materials need OCR at all? Most collectors who want their digital collection to include any semblance of full-text search will require OCR at some point in the digitization process. Without OCR, collections are reduced to images without content. This might be all some collections need, but most projects will benefit from the added layer of text.

Full-text search is dependent on the OCR quality and enhances a user's search accuracy. For instance, someone using the final product might search for "**Tolkien**" to look for the author in the collection. Without the text data overlay, there would be no information in the images to search. But if the document has been properly OCRed, a user could sort individual documents in the collection by author name (quickly leading

them to any works by Tolkien in the collection) or search the entire collection for all mentions of the author within a document.

Intricate metadata and indexing add another level of searchability to the final collection, but these factors don't necessarily rely on accurate OCR. Metadata allows authors, dates, topics and other identifiers within the text to serve as anchoring points for digital archivists to tag parts of the document without performing a line-by-line proof of the text. With metadata alone, search functionality will be limited to the key terms that the archivist creates. Whereas with a complete text layer, any key term can be searched for throughout the document.

Collections that might not benefit from OCR include image-only digital libraries utilizing only metadata and needing limited search functionality. This includes collections that may be on a tighter budget or plan to more closely proof text in the future. However, collectors should keep in mind that it costs less overall to do the job right the first time.

How Does Anderson Archival Do OCR Differently?

Some digitization or imaging companies offer OCR services for a negligible cost. Generally, these companies run an automated OCR program in conjunction with feeder scanning to churn out processed pages and data. While this does dramatically cut down on time spent per page, even the best software needs human guidance for accuracy.

Depending on the quality of the scanned image, the software might read smudged or speckled or skewed characters incorrectly. For example, when reading an image containing text with tight kerning, the software might read every instance of the word "**born**" as "**bom**" in the OCR text file. Using OCR software alone creates inaccuracies in the text, which then leads to errors in the search functionality. Using the Tolkien example above, if during the OCR quality assurance step the archivist missed correcting a typo from "**Tolkicn**" to "**Tolkien**" that instance will not appear in the search results.

Another situation that OCR doesn't handle well is handwritten text. Although most OCR software can easily read multiple kinds of fonts and formats, letters or other documents with cursive script aren't as readable. In these cases, an expert will carefully transcribe the contents of the letter into the OCR file so that the file image can be exported with the text data layer.

Whether the document being OCRed is a clean, typed page or a smudgy, handwritten note, Anderson Archival's digitization experts take extra quality assurance steps throughout the OCR process to check every line, word, and character against the original text; how comprehensive these steps are depends on the needs and preferences of the collector. Digitization plans should be customized, and Anderson Archival takes the OCR process seriously no matter what a collection requires!