IT化スモールスタート解説（第7回）

OCRとは？ペーパーレス化には「OCR」が必要に

2020.03.30

ペーパーレス化のための技術として、真っ先に挙げられるものが「OCR」です。名前は聞いたことがあっても、実際に使ったことはない、具体的にどういうものか分からない、という方も多いのではないでしょうか。

今回はOCRが具体的にどのような技術なのか、利点や注意点、精度を上げるための方法などについて解説します。

OCRとはどのような技術か？

OCRはOptical Character Recognition／Readerの略称で、光学的文字認識のことを表します。紙面の手書き文字や、印刷された文字をスキャナーやカメラなどで読み取り、デジタルデータ化する技術です。

通常、スキャナーやカメラで読み取った紙面は、画像データとして保存されます。しかし、OCR専用ソフトを用いることで、文字を認識しテキストデータとして保存できるのです。アナログデータではなく、デジタルデータとして保存できるため、書類整理や文書入力などに活用されます。

新聞紙などをOCRでデジタルデータ化する例で、もう少し詳しく紹介します。

はじめに、新聞紙をスキャナーなどで読み込むと、それを画像データとして保存します。そして、どの部分が文字の領域なのかを認識するために、レイアウト解析が実施されます。文字の部分と図や表などの画像部分に分け、レイアウト解析で切り出した文字を認識してテキストデータに変換します。最後に元の文書と同じ形で利用できるように、フォーマット出力を行う、というのが一連の流れです。

OCR専用ソフトは、だいたい次のフォーマットに対応しています。

・Word
・Excel
・PDF
など

なお、現在ではOCRとAI（人工知能）を組み合わせた「AI OCR」も登場しています。

OCRの利点

OCRは業務効率化に役立つ技術です。以下からは、OCRによってどのような業務効率化が可能になるのか、具体的に紹介します。

＜ペーパーレス化の促進＞
OCRでは、紙資料のようなアナログデータをデジタルデータに変換します。紙資料は不要となり、廃棄・整理が行えます。紙資料は物理的にスペースを専有します。保管スペースの確保に苦労している方にとっては、大きなメリットといえるでしょう。ダンボール1箱分の書類データ化した場合、数メガバイト（MB）程度で収められる可能性もあります。

さらに、伝票などの情報を手入力でデータ化している場合、作業スピードの向上につながるほか、手入力による入力ミスの回避にもつながります。

＜検索・コピー＆ペースト・編集ができる＞
OCRによってデジタルデータに変換できれば、文書検索やコピー＆ペースト、データの編集ができるようになります。特に、膨大な書類の中から必要な情報を探す場合、文書検索ができ大きな効率化につながるでしょう。コピペや編集も簡単に行えるため、紙資料で保管するよりも管理にかかる手間が大幅に減ります。

OCRの注意点

OCRは非常に便利ですが、いくつか注意が必要です。まず、紙資料から読み取るので、資料の質によっては完璧にスキャンできないときがあります。

次のような場合に、スキャンが失敗する可能性があるので気を付けましょう。

・カラー文字や文字のかすれ
・文章が斜めに印字されている
・文字の間隔が詰まり過ぎている
・特殊文字「㎡（平方メートル）」や「™（商標）」などが使われている
・手書きのクセのある文字
・縦書きと横書きが混在している

上記に該当する場合は、OCRが紙面の文章をテキストデータに変換するのが難しく、スキャンが失敗しがちです。しっかりとスキャンできているかを目で確認しなければなりません。

また、スキャナーなどでスキャンする際には「読み取り解像度」にも注意が必要です。読み取り解像度は、どれだけ細かく読み取りを行うかの設定となります。読み取り解像度が低いと、OCRはうまく文字を認識しません。

読み取り解像度は、200～300dpi程度の解像度でスキャンしましょう（dpiとは「dots per inch」の略称で、1インチあたりにどれだけのドットを表現できるかを示す「ドット密度」の単位。数値が大きいほど読み取り解像度は高くなる）。

その他にも、OCRでスキャン精度を上げる方法もあります。次で詳しく紹介します。

OCRの精度を上げるための方法

OCRの精度を上げるには、上で挙げた「OCRの注意点」に気を付けながら、次の方法を実施しましょう。

＜傾きを補正する＞
紙資料をデータ化した場合に斜めにスキャンしてしまったら、スキャンし直すか、画像ソフトを使って傾きを補正しましょう。

＜カラー原稿を白黒でスキャンする＞
OCRは、カラー原稿よりも白黒原稿のほうが認識率は向上します。カラー原稿でも高い認識率を保つOCRソフトが登場していますが、白黒でデータ化すると、より精度を高められます。

＜コントラストを強調する＞
コントラストを強調して文字を濃くすると、ハッキリとスキャンできます。スキャナーの機能には、コントラスト調整があります。読み取り時に試してみてください。

＜裏写りを軽減する＞
裏写りすると、余計な情報がデータに反映されることになります。裏写りの軽減を意識しましょう。裏面が写り込んでしまう薄い紙資料の場合は、裏写りしないように黒紙を下に敷くなどの対処をしましょう。

OCRでペーパーレス化を進めよう！

OCRは、スキャナーなどで紙面の文字情報を読み取り、テキストデータに変換する光学文字認識技術です。スキャナーなどで読み取ったものは画像データとして保存されますが、OCR専用ソフトを利用すれば、文字の部分をテキストデータとして保存できます。

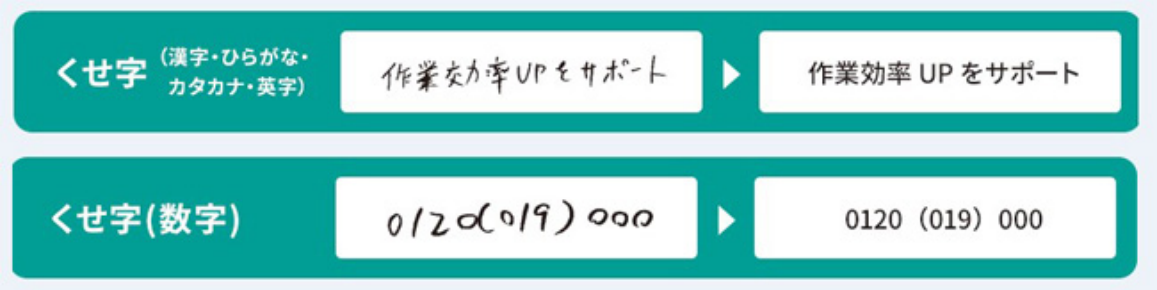
OCRを用いれば、大量の紙資料をデジタルデータに変換できます。データ化で、紙資料の保管スペースを有効活用したり、情報を探す手間を減らして業務効率化につなげたりすることが可能です。

ただしOCRを利用する際には、読み取り精度を妨げるいくつかの注意点に気を付けましょう。

紙媒体を利用した業務は非常に手間がかかるものです。紙資料から手入力でデータ化しようとすると入力ミスの可能性もあるため、OCRを使って効率的に業務を進めるのをお勧めします。

NTT西日本では、AIを用いたOCRサービス「おまかせAI OCR」を提供しています。データ入力に割いていた業務時間の短縮化や、書類削減による省スペース化など、業務全体の効率化に貢献するサービスです。

従来は読み取りが難しかった「手書きのクセ字」もAIを用いることで、高精度な文字認識が可能となりました。手書き訂正や訂正印のある紙資料も判別可能です。



専用のサポートセンターを用意し、「操作方法のご案内」や「トレーニング」からサポートします。もちろん、困りごとのフォローもお任せください。

働き方改革の対応の一環として、業務効率化を図りたい方や手書き入力のデータ化などでお困りの方は、ぜひ気軽にお問い合わせください。

※「おまかせAI OCR」は、AI inside 株式会社の「DX Suite」を活用しています
※「DX Suite」はAI inside株式会社の登録商標です

※掲載している情報は、記事執筆時点のものです