OCRリーダーのご提案

今後いっそう二次利用の必要性が高まる書籍の文字情報。
紙媒体からのテキストデータ抽出は明昌堂のOCRシステムをご活用ください。

ほとんどの書籍においてその中核をなし、基本となるデータ要素といえば、やはり文字情報です。

文字情報は、現在のDTPフローでは原則的にテキストデータとして扱われます。当然、テキストデータ化された文字情報は、書籍が完成した後でも、改訂新版を制作する、他の書籍へ記事を転用する、Webへ掲載するなど、さまざまな流用が可能です。特に近年では電子書籍化の目的でテキストデータの必要性が高まっています。

しかし、書籍のテキストデータが適切に管理されていることはまれで、多くは二次利用の際にDTPデータから取り出すことになります。一方、DTPデータが保管されていない、そもそもDTP制作していないなどの理由から、原稿として使用できるのは、書籍（紙媒体）のみというケースも多くあります。

当社では、紙媒体に印刷された活字も二次利用に活かせる文字情報であると考えており、紙媒体の文字情報をPCで使用できるテキストデータに変換する手段として、OCRを活用しています。

■OCRリーダーについて

OCRリーダーとは紙媒体などをスキャンして画像化し、そこから文字情報を自動的に読み取って、テキストデータ化するソフトのことをいいます。

過去のOCRリーダーは性能が低く、データ化した後の整理・修正に逆に手間がかかってしまうことも多かったのですが、近年は識字エンジンの能力が飛躍的に向上し、当社導入ソフトにおいて99.8%の識字率（当社試験による）という結果が出ており、十分実用に耐える性能となっています。

図1　連続自動処理フローの概要。監視フォルダに入った画像は自動的に処理されます

■当社のOCRシステム

当社ではメディアドライブ社の「WinReader PRO v.13.0」を導入し、監視フォルダからの自動テキスト化フローを構築しています（図1）。

スキャナから取り込んだ画像が監視フォルダに入ることで、縦組・横組・表組などのレイアウトを自動で分析し、識字作業を開始、出力フォルダにテキストが生成されます。

紙面構成が複雑な書籍の場合はレイアウトを手動で選択することで、識字率を下げずに処理することも可能です（図2）。

■OCRの特性

ソフトの性質から、図表が少なく文字がメインのものが最も作業効率が高くなります。ルビには未対応です。また、コピー機で複写された原稿などは若干識字率が落ちます。和欧文の混在が激しいことも識字率の低下の原因になります。

■品質の保持対策

当社では生成したテキストを校正ソフト「Just Right!」にかけ、OCR特有の促音の誤認識などを検査しています。その後プリントアウトしたテキストを人間の目でもう一度原稿と付け合わせて校正する二重のチェックを行っています。

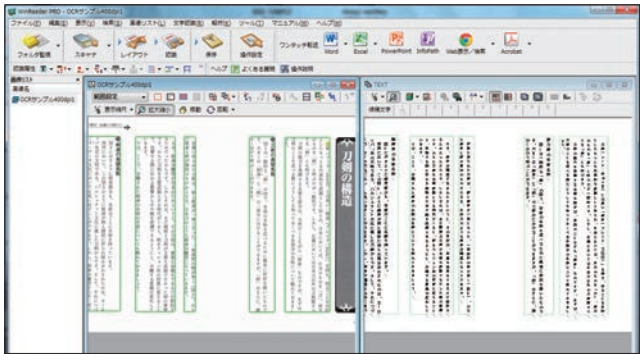

図2　左のウインドウでスキャン画像の変換範囲を個別指定することができます。右のウインドウには変換されたテキストが表示されています