数式を含む文書(紙、PDF)をWordデータに変換する

InDesignなどで作成されたDTPデータ、または紙の文書を
専門的なOCRソフトを活用してWordデータに変換する方法をご紹介します。

■組版データのWord変換

Wordは最大シェアのテキスト作成アプリケーションです。誰でも簡単に扱うことができ、DTPのテキスト原稿はほとんどがWord形式で入稿されます。

一方、できあがったDTPデータをあらためてWordへ変換してほしいという要望が、特に教科書、教材、学術書を制作されているお客様より以前から多く寄せられていました。Wordデータにすることで手元で編集し、二次利用するのがその主な目的です。

これまでDTPデータをWordに変換する一番効率的な方法と考えられていたのは、まずDTPデータからPDFを書き出してそれをWordで開き、Word上で体裁を整えるという工程でした。当社のMCR Vol.45およびVol.70でも詳しく紹介しています。

しかし、この方法では数式などの複雑な文字組みは再現できず、体裁を整えるのに一からWordで組版するのと変わらないほどの手間がかかることがあります。

そのため当社ではよりよい変換方法を探してきましたが、専門的なOCRソフトを活用することにより、今まで以上に効率良く、また数式を含む組版データなども正確に変換できることが分かりました。

■InftyReaderの活用

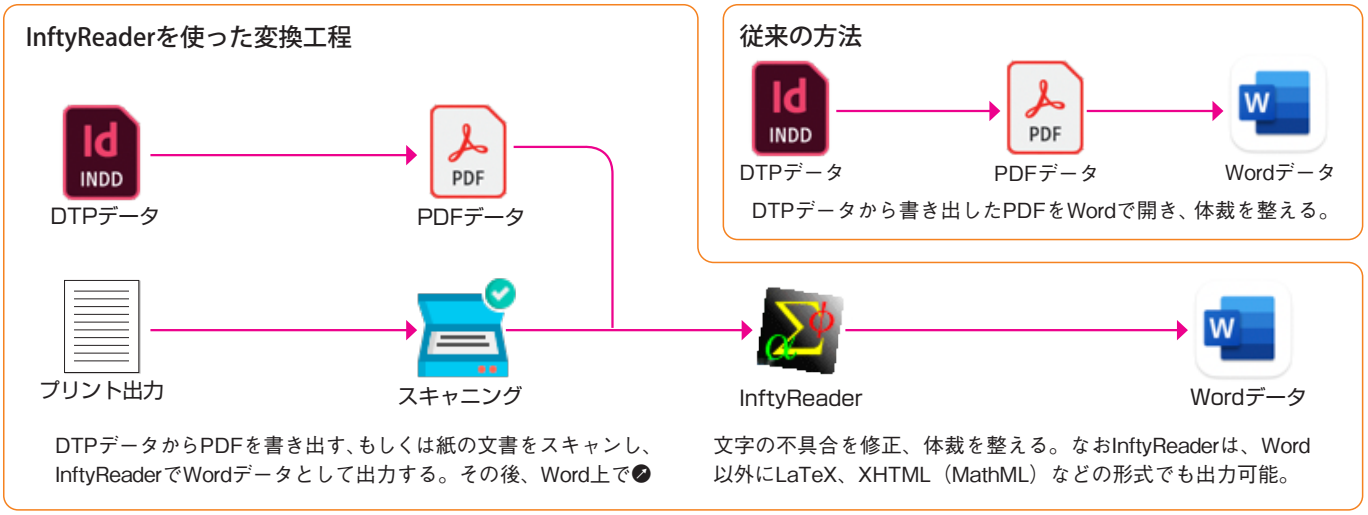
今回紹介します「InftyReader」は、数式を含む文書を処理できるOCRソフトです。

InftyReaderの変換工程は下図のようになっています。変換元の文書はフォーマットを問わず、紙の文書からでも変換が可能です。実際の変換結果はウラ面をご覧ください。

文字スタイル､段落スタイルは移行できませんので、Word上で調整する必要があります。また、表組みや図版が混在していると正しく変換できません｡しかし、数式部分は非常に高い精度でWordの数式エディタに変換されます。

シンプルなレイアウトの紙面であれば､InftyReaderでかなり正確な文書変換が行えます。複雑なレイアウト紙面については、例えば元の文書を適宜トリミングし、再整列するなどの事前調整を行うことによって、精度の高い変換結果が得られます。

当社では出版印刷業界内で認知度の低いソフトウェアなども活用して、データ制作の問題解決に取り組んでいます。組版データの加工や再利用について要望がありましたら、是非ご相談ください。



InftyReaderの数式変換

入力ファイル（PDF）

1 次の問いに答えなさい。

(1) 次の計算をしなさい。

① $-5-(-12)$

② $35\div(-3)^2$

③ $\frac{x+y}{5}+\frac{2x-y}{7}$

④ $\sqrt{3}\,(3\sqrt{4}-\sqrt{12})$

出力ファイル（Word）

1
次の問いに答えなさい。
(1)
次の計算をしなさい。
①
$-5-(-12)$
②
$35\div(-3)^2$
③
$\frac{x+y}{5}+\frac{2x-y}{7}$
④
$\sqrt{3}\left(3\sqrt{4}-\sqrt{12}\right)$

数式や分数を含む文書を変換した。字下げなどレイアウトの再現性はないが文字認識の精度は高い。数式部分はWordの数式機能に変換されている。

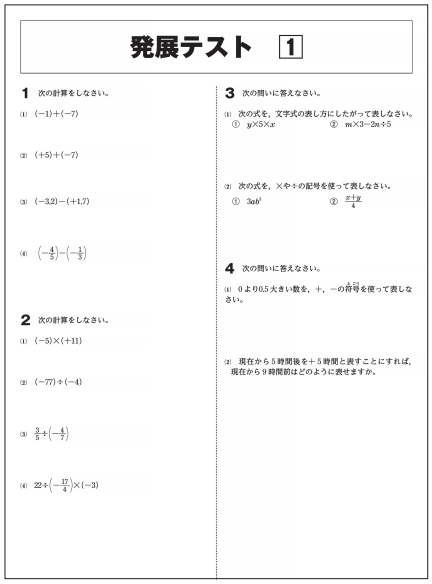
このような数式部分はWordの数式機能で再現される。

分数混じりの数式も問題なく再現される。

出力ファイル(Word)から書き出したPDFをご覧いただけます。

調整サンプル①

元の誌面


発展テスト 1

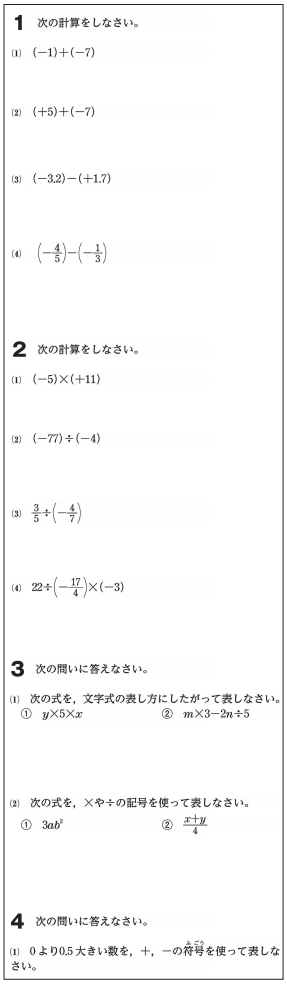
1 次の計算をしなさい。
(1) $(-1)+(-7)$
(2) $(+5)+(-7)$
(3) $(-3.2)-(+1.7)$
(4) $\left(-\frac{4}{5}\right)-\left(-\frac{1}{3}\right)$

2 次の計算をしなさい。
(1) $(-5)\times(+11)$
(2) $(-77)\div(-4)$
(3) $\frac{3}{5}\div\left(-\frac{4}{7}\right)$
(4) $22\div\left(-\frac{17}{4}\right)\times(-3)$

3 次の問いに答えなさい。
(1) 次の式を，文字式の表し方にしたがって表しなさい。
① $y\times5\times x$　② $m\times3-2n\div5$
(2) 次の式を，×や÷の記号を使って表しなさい。
① $3ab^2$　② $\frac{x+y}{4}$

4 次の問いに答えなさい。
(1) 0より0.5大きい数を，＋，－の符号を使って表しなさい。
(2) 現在から5時間後を＋5時間と表すことにすれば，現在から9時間前はどのように表せますか。

入力ファイル(PDF)


1 次の計算をしなさい。
(1) $(-1)+(-7)$
(2) $(+5)+(-7)$
(3) $(-3.2)-(+1.7)$
(4) $\left(-\frac{4}{5}\right)-\left(-\frac{1}{3}\right)$

2 次の計算をしなさい。
(1) $(-5)\times(+11)$
(2) $(-77)\div(-4)$
(3) $\frac{3}{5}\div\left(-\frac{4}{7}\right)$
(4) $22\div\left(-\frac{17}{4}\right)\times(-3)$

3 次の問いに答えなさい。
(1) 次の式を，文字式の表し方にしたがって表しなさい。
① $y\times5\times x$　② $m\times3-2n\div5$
(2) 次の式を，×や÷の記号を使って表しなさい。
① $3ab^2$　② $\frac{x+y}{4}$

4 次の問いに答えなさい。
(1) 0より0.5大きい数を，＋，－の符号を使って表しなさい。

2段組レイアウトのまま文字認識を行うと誤変換や文字飛びが発生した。そのため右図のように、2段組の誌面を切り分けて上下に繋げる調整を行ってから変換を行った。

調整前

調整後

調整サンプル②

入力ファイル(PDF)

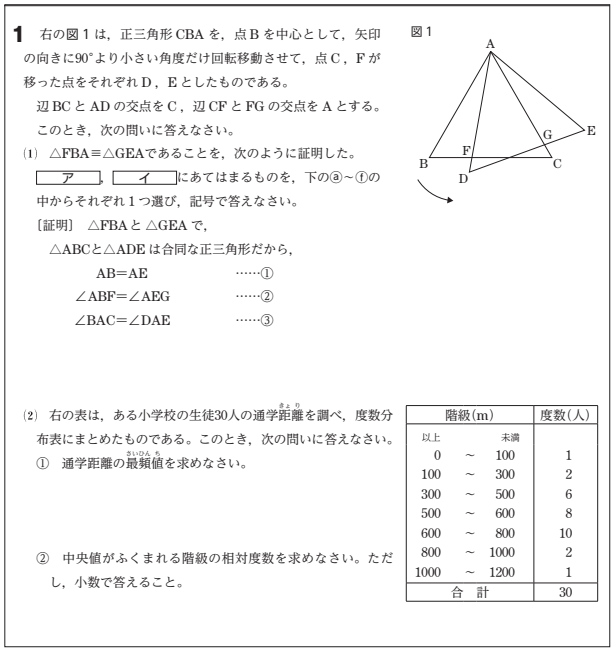

1 右の図1は，正三角形CBAを，点Bを中心として，矢印の向きに90°より小さい角度だけ回転移動させて，点C，Fが移った点をそれぞれD，Eとしたものである。
辺BCとADの交点をC，辺CFとFGの交点をAとする。
このとき，次の問いに答えなさい。

図1

(1) △FBA≡△GEAであることを，次のように証明した。
［ア］，［イ］にあてはまるものを，下のⓐ～①の中からそれぞれ1つ選び，記号で答えなさい。
［証明］ △FBAと△GEAで，
△ABCと△ADEは合同な正三角形だから，
AB=AE ……①
∠ABF=∠AEG ……②
∠BAC=∠DAE ……③

(2) 右の表は，ある小学校の生徒30人の通学距離を調べ，度数分布表にまとめたものである。このとき，次の問いに答えなさい。
① 通学距離の最頻値を求めなさい。
② 中央値がふくまれる階級の相対度数を求めなさい。ただし，小数で答えること。

階級(m)			度数(人)
以上		未満	
0	～	100	1
100	～	300	2
300	～	500	6
500	～	600	8
600	～	800	10
800	～	1000	2
1000	～	1200	1
合　計			30

表組みや図版が混在しているとそれらに含まれている文字が読み取られ、予期せぬ場所に挿入されてしまう。表組みや図版を削除してから変換することにより、正確な変換結果が得られる。

調整前

調整後