

Extract.exe

概要

pdf ファイルから、画像、テキストを切り出す。

コマンドライン

Extract [ファイル名]* [オプション]*

指定のファイルを順番に読み込み、-D または、-C オプションで指定したフォルダーにテキストファイル、あるいはグラフィックファイルとして書き込みます。

また、出力ファイルのリストを格納したリストファイルも書き込みます。

出力ファイルの命名規則

テキストは、ブロック毎に別のファイルとして出力します。グラフィックはそれぞれの画像を単独のファイルとして出力します。

ファイル名は、prefix ページ番号_連番.拡張子の形式です。ページ番号、連番はそれぞれ 1 から始まります。テキストの場合、拡張子は txt、グラフィックの場合は、それぞれの画像の内部形式に従います (jpg、png など)。

prefix はデフォルトでは空ですが、固定の文字の指定 (-P)、ファイルのベース名 (-QF)、ファイル連番 (-QN) を指定できます。

リストファイルは、標準ではテキスト、イメージに対してそれぞれ、prefix textlist.txt、prefix imagelist.txt が対応します。オプション (-U) を指定するとリストファイルは統合され、prefix listfile.txt を出力します。

オプション

-Cfoldername

出力フォルダを指定します。フォルダが存在していなければ作成します。フォルダがある場合には、そのフォルダ直下のファイルをすべて削除します。

-Dfoldername

出力フォルダを指定します。既存のフォルダを指定してください。

-I

イメージファイルだけを出力します。

-M

モニタ出力をします（通常は使用しません）。

-Pprefix

出力に固定の prefix を付けるときに指定します。

-Qoption

プレフィックスの形式を指定します。

-Q 固定プレフィックス（通常指定は不要）

-QF ファイルのベース名+「_」をプレフィックスにします。

-QN 1から始まる連番をプレフィックスにします（8桁固定）。

-Roption

テキストのブロック化で使用する広さを指定します。

-RN2 狭い

-RN1 やや狭い

-RS 表示運

-RW1 やや広い

-RW2 広い

-T

テキストだけを出力します。イメージファイルは出力しません。

-U

テキスト、イメージのファイルリストを一つのリストで出力します。

-V

テキストの行内の順序を座標を使用せずに記述順とします。

-X

テキストの横座表を逆順とします。