2022/04/14
有限会社サイパック

PDF に変換すると康熙部首に置き換わる問題について

Unicode のブロックの一つである康熙部首（214 部首）にも使われている文字（常用漢字、第一水準～第三水準の文字）が含まれる文書を PDF に変換した後、テキストを抽出すると当該文字が元の通常の漢字ではなく康熙部首に置き換わってしまう問題があります。以下は、Word のファイルを PDF へ変換した場合の康熙部首への置き換わりを確認した資料です。

■ PDF の作成

康熙部首（214 部首）とそのコードおよび一般の文字コードを記載した一覧表を Word で作成し、その表を下記の 4 フォントでそれぞれレイアウトした上で下記の 4 パターンで PDF へ変換した。

1. ファイルメニュー→名前を付けて保存（別名保存）
2. ファイルメニュー→印刷→「Microsoft print to PDF」を選択して保存（MS print）
3. ファイルメニュー→エクスポートを選択して保存（エクスポート）
4. Adobe Acrobat から PDF を作成（Acrobat）

フォント

a. メイリオ
b. UD デジタル教科書体(Windows10 から追加されたフォント、今回は N-B を使用)
c. Yu Gothic UI(Windows10 から追加されたフォント)
d. MS 明朝

■ 確認内容と結果

変換した PDF のテキスト部分をコピー&ペーストして Word へ貼り付け、変換された文字の右側にカーソルを置き Alt+X キーで文字コードを表示して結果を確認した。変換結果詳細はフォント単位で記載した別途資料「康熙部首（Kangxi Radicals）問題の調査／PDF 変換テスト結果一覧」を参照。

1. 名前を付けて保存
   - a. メイリオ　康熙部首に置き換わった文字はなかった。
   - b. UD デジタル教科書体 N-B　康熙部首に置き換わった文字はなかった。
   - c. Yu Gothic UI　康熙部首に置き換わった文字はなかった。
   - d. MS 明朝　康熙部首に置き換わった文字はなかった。

2. Microsoft print to PDF
   - a. メイリオ　「尢」「幺」「長」「鬼」のみ康熙部首へ置き換わっていた。
   - b. UD デジタル教科書体 N-B　「彐」「疒」「冂」「耂」「黄」以外の文字は画像化されていた。またこれらについて康熙部首に置き換わった文字はなかった。
   - c. Yu Gothic UI　「尢」「幺」「長」「鬼」のみ康熙部首へ置き換わっていた。
   - d. MS 明朝　康熙部首に置き換わった文字はなかった。

3. エクスポート
   - a. メイリオ　康熙部首に置き換わった文字はなかった。
   - b. UD デジタル教科書体 N-B　康熙部首に置き換わった文字はなかった。
   - c. Yu Gothic UI　康熙部首に置き換わった文字はなかった。
   - d. MS 明朝　康熙部首に置き換わった文字はなかった。

4. Acrobat
   - a. メイリオ　「分」を除くほぼ全ての文字が康熙部首へ置き換わっていた。また「尢」「幺」「長」「鬼」については康熙部首とは異なるコードに変換されていた。
   - b. UD デジタル教科書体 N-B　康熙部首に置き換わった文字はなかった
   - c. Yu Gothic UI　「分」を除くほぼ全ての文字が康熙部首へ置き換わっていた。また「尢」「幺」「長」「鬼」については康熙部首とは異なるコードに変換されていた。
   - d. MS 明朝　康熙部首に置き換わった文字はなかった。

以上