OCR Server (Optical Character Recognition)

V ÚCHP byl pořízen rozšiřující OCR systém pro barevnou kopírku v knihovně, převádějící naskenovaný dokument (ve formě bitmapového obrázku) do editovatelného textu ve vybraných formátech. V současnosti máme zvoleny výstupní formáty PDF, XLS, DOC, TXT a ePub. Nyní tedy v naskenovaných PDF dokumentech lze prohledávat text, čísla a některé speciální znaky. Požádejte obsluhu kopírovacího stroje o dodání PDF dokumentů, zpracovaných pomocí OCR systému.

Pokud používáte svůj stolní skener, máte již mnoho naskenovaných dokumentů v PDF bez zpracování pomocí OCR, případně používáte mobil nebo tablet jako rychlý skener, můžete nechat zpracovat dokumenty dle níže uvedeného postupu a využít výhody tvorby svého digitálního archivu, ve kterém lze text vyhledávat.

Postup zpracování osobních skenů:

1. Připojte si na svém PC s Windows adresář \\ocrserver.asuch.cas.cz\HotFolder (v Linuxu nebo Mac OSX pak smb://ocrserver.asuch.cas.cz/HotFolder)
2. V připojeném adresáři si zvolíte vstupní adresář dle požadovaného výstupního formátu. `sPDF, excel, word, txt, epub`
3. Do zvoleného vstupního adresáře zkopírujete svůj naskenovaný dokument. Povolené formáty jsou PDF, JPG, TIF, GIF, PNG.
4. Ve výstupním adresáři *Out najdete zpracované soubory. Zkopírujte si je na svůj disk a vymažte na serveru.

Detailní postup zpracování osobních skenů:

1. Připojte si na svém PC s Windows adresář \\ocrserver.asuch.cas.cz\HotFolder (v Linuxu nebo Mac OSX pak smb://ocrserver.asuch.cas.cz/HotFolder)
2. V připojeném adresáři si zvolíte vstupní adresář dle požadovaného výstupního formátu. `PDF - sPDF, XLS - excel, DOC - word, TXT- txt, ePub - epub`
3. Do zvoleného vstupního adresáře zkopírujete svůj naskenovaný dokument. Povolené formáty jsou PDF, JPG, TIF, GIF, PNG.
4. Za malý okamžik zkopírovaný soubor ze vstupního adresáře zmizí, protože se dostal ke zpracování. Vlastní proces OCR trvá určitou dobu, závislou na počtu stránek vstupního souboru. Typicky zpracování 1 strany A4 převáděné do PDF trvá 30 vteřin. Výsledek po zpracování celého souboru se objeví v sousedním adresáři s příponou `Out`. Např. u PDF je vstupním adresářem `/sPdf` a výstupním `/sPdfOut`
5. Ve výstupním adresáři najdete 2 soubory. Jednak zpracovaný soubor s příponou dle zvoleného formátu např. PDF a druhý s příponou `.result.xml`, jenž obsahuje technické údaje o průběhu procesu. Můžete si jej prohlédnout ve webovém prohlížeči. Na konci tohoto souboru je užitečná informace o výsledku zpracování např. `<Statistics TotalCharacters=„910“ UncertainCharacters=„49“ PagesArea=„1“`. Po překopírování souborů na svůj disk, soubory na serveru vymažte.

Technické informace

- Použitý software: ABBYY Recognition Server 3.5
- Doporučené rozlišení skenu: 200 dpi (u textů s malou velikostí 300 dpi)
- Podporované jazyky slovníků pro lepší výsledky OCR: čeština, angličtina, němčina, francouzština, španělština (Systém podporuje až 160 jazyků)
- Licencovaná kapacita systému: 5000 stránek/kalendářní měsíc (lze rozšířit)

V případě nejasností nebo problémů volejte výpočetní středisko