Formát PDF je výborný pro archivaci dat, faktur apod. Takový soubor si přečtete téměř na libovolném operačním systému a zařízení, včetně elektronických čteček či mobilních telefonů s Androidem. Pokud na počítači s Windows nechcete instalovat tak trochu megalomanský Adobe Reader, vyzkoušejte kupříkladu (portable) aplikaci Sumatra PDF.

Dříve jsme byli nuceni export z Office provádět prostřednictvím tisku na virtuální PDF tiskárnu (z instalace GhostScriptu, velmi dobrý BullZip s podporou pro VBA a příkazový řádek aj.). Od verze Office 2007 (se service packem) je již export do PDF dostupný v rámci Excelu nativně. Nedej bože, pokud ale z PDF potřebujeme dostat data zpět do Office. Word 2013 sice přichází s editací takových dokumentů, ale…

Mějme soubor, ze kterého jsem čerpal data obsahující informace o množství vitaminu C v potravinách. Jedná se o dvě dvousloupcové tabulky vedle sebe (potravina – obsah vitaminu C).

Běžný výběr textu v PDF

1. Text se pokusím vybrat v Readeru běžným způsobem, zkopírovat přes schránku a vložit do Excelu (ale i jinam). Výsledek nebude valný.

kapusta růžičková 787 melouny, dýně 220
kedlubny bílé 448 rybíz červený, bílý 330
křen 1 125 rybíz černý 1 360
květák 383 ananas 206

Nejenže Excel nepochopí obsah jako tabulku, ve výsledku nejsou rozlišeny hranice buněk (mezi texty, čísly i jako oddělovač tisíců jsou prosté mezery). Poradíte si s tím v Excelu? Horko těžko. Podle mezer můžete text naporcovat – na listu přes Data / Text do sloupců, pod VBA přes pole, Split a dvojité užití WorksheetFunction.Transpose. V obou případech si pak ale budete muset najít čísla a vracet mezery tam, kam patří.

Leckdy špatný formát PDF způsobí při kopírování přes schránku degradaci českého kódování a podle mě není cesty, jak si s tím poradit. Je potřeba také zdůraznit, že formát PDF umožňuje ochranu proti kopírování a tisku. A ano, existují nástroje, které ji odstraní, ale z pochopitelných důvodů zde nebudu psát návod, jak to provést.

Tip: Víte o tom, že pokud v řadě textových editorů použijete klávesu ALT před samotným textovým výběrem, probíhá výběr přes sloupce a ne řádky? A v Readeru to také jde (klávesa ALT musí být držena ještě před označením počátku tažení myškou!). Není to zázrak, ale při psaní článku jsem postupoval stejně.

Výběr textu v PDF s přidržením klávesy ALT (stav před)

Výběr textu v PDF s přidržením klávesy ALT (stav po)

2. Adobe Reader umožňuje exportovat z PDF text (Soubor / Uložit jako jiné / Text). Výsledek nebude použitelný, stejně jako v předchozím způsobu.

3. Adode Reader obsahuje volbu Soubor / Uložit jako jiné / Word nebo Excel online (odkaz). Je to ale jen lákadlo, které vás přijde na cca 25 eur ročně a nevím, nakolik je tento nástroj kvalitní.

4. Nevím o freeware nebo levném editoru, který by daný problém řešil (např. Foxit PDF Editor). A za verzi Adobe Acrobat XI Standard, který by to měla umět, zaplatíte 138 eur…

5. Mám hodně rád HyperSnap pro snímání obrazovky, který uměl i přebírat z okna text (např. seznam souborů z okna Total Commanderu). Naneštěstí leckdy nečisté technologie vykreslování, DirectX apod. tento nástroj poslaly k ledu a tak jeho tvůrci od něj prakticky upustili. Jednoduše řečeno, HyperSnap není schopen rozpoznat text v okně Adobe Readeru. Okna už jednoduše nejsou, co bývala, a nejspíš by to bylo peklo i s podporou API.

6. Řešení nabízí regulární výrazy. Pro práci s nimi mně osobně vyhovuje RegexBuddy (který ale stojí 30 eur). Pokud potřebujete freeware, zkuste se podívat na Google.

Regulární výrazy – analýza textového řetězce (oddělovač svislice)

Regulární výrazy – analýza textového řetězce (oddělovač tabulátor)

Šablona (maska) v regulárním výrazů říká, že hledám celá čísla, mezi nimiž může a nemusí být mezera a do výběru přidávám i mezery před a za. V rámci nahrazování používám „backreference“ (vnější závorky v šabloně a následně \1, ve VBA níže $1). Kouzlo druhého uvedeného příkladu je v tom, že tabulátor coby oddělovač (\t) Excel pochopí jako hranici buňky a při vkládání výsledku regulárního výrazu ze schránky jednoduše obsah rozdělí do buněk, provede i do jisté míry ořez a pochopí formát. Pozn. Pokud zkopírujete výsledek se svislicí, vložíte jej ze schránky do listu, provedete rozdělení prostřednictvím Data / Text do sloupců, tak při opakovaném kopírování ze schránky bude Excel svislici již brát jako hranici buňky bez nutnosti dalšího zpracování! Takové chování se promítá i kupříkladu při načítání CSV souborů do Excelu.

Tip: O regulární výrazy můžete obohatit i VBA (objekt RegExp je obsažen ve VBScriptu, WSH), bohužel se jedná o starší verzi, která ne úplně dobře respektuje českou znakovou sadu, neumí některé dopředné a zpětné vyhledávání a zástupné symboly.

Regulární výrazy pod VBA

Uvedené zpracování ve VBA mělo ještě kromě popisovaných problémů další háček. Kopírováním z okna Immediate přes schránku jsem přišel o správné kódování… A pak najednou z ničeho nic daný problém vyšuměl. Je to alchymie… K regulárním výrazům se určitě někdy vrátíme.

7. Svého času jsem prováděl testování specializovaných aplikací typu „pdf to xl“, ať už těch desktopových, nebo online. A i dnes je mým vítězem PDF2XL. Trial verze vám poslouží na 7 dní a 50 konverzí, do výsledku přidá dodatečné informace, které ale nečuní kopírovaná data. Placená verze PDF2XL Basic 5 přijde na docela dost – 82 eur. Ostatní aplikace vykazovaly řadu chyb – od kódování češtiny po špatně naformátované buňky po překopírování.

Zde je ukázka podrobnějšího testu v PDF2XL. Překonatelným problémem byla nutnost korekce hlavičky a prvního řádku (nutný ruční split).

PDF2XL – detailnější testování

Na rozdíl od dříve testované verze 3 jsem nemusel ve verzi 5 už zasahovat do nastavení formátu

buněk, ale pokud potřebujete, navrhuji následující úpravy:

PDF2XL – nastavení

8. Vždycky mějte na paměti jednu věc – co vidíte, dokážete i zkopírovat. To platí pro elektronickou i papírovou formu dat, tabulek, fotografií. Jak? Kouzlo se skrývá za skenováním (u papírové dokumentace) a třemi písmenky – OCR, čili rozpoznáváním textu. Podle mě nejsofistikovanějším softwarem pro tyto účely je ABBYY FineReader, který si velmi solidně poradí i s češtinou a tabulkami (bohužel nemohu v tuto chvíli nabídnout relevantní obrázek).

Výsledky kopírování do Excelu:

Výsledky kopírování z PDF

PDF2XL – výsledek podrobnějšího testu

Testovací soubory:
excel-test-pdf.zip