Zpravodaj Československého sdružení uživatelů TeXu

Pavel Stříž; Radek Benda

Editace PDF souboru aneb O jednom dnu

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 20 (2010), No. 1-2, 14–22

Persistent URL: http://dml.cz/dmlcz/149991

Editace PDF souboru aneb O jednom dnu

Pavel Stříž, Radek Benda

Abstrakt

Tento článek popisuje experimenty nad editací a hromadným nahrazováním výseků textů v PDF souborech. Tento příspěvek není pro slabší povahy, protože většina experimentů byla v principu neúspěšná a nevedla ke kýženému výsledku. Jsou také prezentovány otevřené myšlenky a úvahy jednotlivých řešitelů. Článek je psán volnější formou v první osobě jednotného čísla.

Klíčová slova: editace PDF, nahrazování textů, Adobe Acrobat, CAD-KAS PDF Editor 2.6, PDFedit, balíček pdfpages.

doi: 10.5300/2010-1-2/14

Předmluva

V tomto článku a v případě čtenářského zájmu v několika dalších článcích se pokusíme spíše volnější formou představit zapeklité typografické situace, které byly řešeny pod časovým tlakem během dvaceti čtyř hodin.

Úvod

„Potřebuji tvou pomoc s bělením!“ vystartoval na mě jeden známý hned poté, co jsem přišel do práce. Že prý to hoří, je toho přes 600 stránek a musí to být do zítřka atd. Já že tedy ano. Začalo mi však vrtat hlavou, o co vlastně jde.

Korekturní bělení jsem přestal používat už na střední škole, pokud jsem ho kdy řádně použil. Celá procedura je náročná. Zabělíte text, necháte vrstvu bělidla vyschnout a uděláte fotokopii. Kvůli vrstvě bělidla nelze použít podavač, takže se musí dělat strana za stranou. Ve většině případů lze poznat, že se jedná o fotokopii. O tiskové kvalitě 600 dpi tedy žádná řeč!

Z kolegy postupně vylezlo, že má promazat určité partie textu proto, že nemůže zasáhnout do sestavy a zpětně ji datovat. Problematika zpětného datování je má oblíbená. Už na základní škole jsem se se spolužáky naučil měnit systémový čas a založit soubor tak, aby vypadal, že vznikl před několika lety. V principu stačí jen vykopírovat obsah na úrovni znaků v šestnáctkové soustavě. Naše oblíbená činnost byla úprava textů ve hrách editací `exe` či `com` souborů. Co se v mládí naučíš… Tohle však byla jiná situace. Jednalo se o tiskové podklady do archivu.

Na vysoké škole si můžeme představit tisk zkouškových katalogů. Za tři roky po ukončené zkoušce sestava může vypsat, že je student již absolvent, což před třemi roky nebyla pravda. Pokud nemáme přístup do sestavy, resp. databázového serveru, jsme bezmocní. Téměř bezmocní.

Z kolegy jsem postupně dostal, že řeší podobnou situaci. Může obsluhovat server jen z předdefinovaného uživatelského hlediska. Při bližším zkoumání u daného počítače jsem zjistil, že si připravuje tisky. Kde lze tisknout, lze použít tiskový ovladač. Obratem jsem si nainstaloval tiskový ovladač pdfEdit995 (`http://www.pdfedit995.com/`), podobně bylo možné zkusit PDFCreator (`http://sf.net/projects/pdfcreator/`) i jiné. Mohl jsem užít i Adobe Acrobat, který je na fakultě zakoupený.

Tedy, to se mi ulevilo, takže fyzické zásahy do papíru nebudou nutné, existuje totiž elektronická verze, dokonce ve vektorové formě. Tím odpadá proces skenování, případného užití OCR programů, úpravy a vlastního tisku. V této chvíli jsem byl již rozhodnut korekturní bělení zamítat, jak to jen šlo, i přes přemlouvání a další naléhání kolegy. Byl by to zabitý, probělený víkend.

V této chvíli kolega zatlačil, jestli se dá stihnout zabělit cca 5000 slov na 600 stranách A4. Já mu oznámil, že lze přímo editovat PS nebo PDF soubor, i když to není typické, že to nejsou formáty určené k editaci. Že to půjde snadno, a že to do zítřka krásně stihnu. Kolegovi spadl balvan ze srdce, když jsem jej na sto procent ujistil, že to zmáknu. Během dne jsem zalitoval svého bláhového výroku, ale jak to v životních příbězích bývá, vše nakonec dobře dopadlo.

Sedíce u programu

Kolega najednou zmizel, hodil to úspěšně na mě a šel si dělat něco svého. Kamarád! Zůstal jsem u magického počítače s přístupem do databáze. Zjistil jsem další zajímavé úkazy. Nejen, že mohu získat PostScriptovou formu, ale lze získat i původní data do libovolného formátu, na který si vzpomenete: TXT, DAT, CSV, XML či HTML.

Otevřel se úplně nový svět řešení tohoto problému. Než však kolega odběhl, předal mi ukázkové tisky, a že to musí splňovat přesně tento grafický návrh. Při eufórii možných exportů jsem na tuto drobnost úplně zapomněl, teď se vrátila.

Situace se nám zkomplikovala. Především v té chvíli, kdy jsem zjistil, že ani při nejlepší vůli nejsem schopen dostat použitelnou PostScriptovou podobu. Přímý výstup nevhodně zobrazoval znaky s diakritikou a podezřele tyto znaky nahrazoval. Při bezproblémovém výstupu do PDF a exportu do PS jsem nebyl schopen jej přemluvit k vektorové formě. Navíc se nenechal přemluvit k zápisu do PS, ale jen do jednostránkového EPS. Takže i kdybychom nakrásně použili spojování PS souborů, tak bychom museli exportovat databázi na šestsetkrát.

Cvičení: Spojte si všechny EPS/PS soubory setříděné abecedně z podadresáře do jednolitého PS souboru.

Shrneme možnosti, které mě tehdy napadly:

- Vzít bitmapovou podobu a PHP a GD knihovnu. Pak by stačilo vyhledat vzor slova, které se má vymazat a nahradit sérii barevných pixelů bílou barvou. Když pomineme neuvěřitelné paměťové nároky (A4 při, řekněme, minimálních 300 dpi), tak tu byl jiný problém. Mazaná slova se pokaždé vyskytla na jiné pozici. To znamená, že v rámci bitmapy ta samá slova nemusí být identická po konverzi z vektorové formy. Tím tato možnost odpadla. Tato možnost by odpadla okamžitě, kdybychom dělali skenování z papíru, tam by navíc byl problém s pootočením objektů a různými nečistotami a stupni šedi.
- Další možnost byla promazat slova na úrovni HTML a dopracovat styl užitím kaskádových stylů (CSS). To nebylo hraní pro mě. Protože navíc musel být zachován stránkový zlom, bylo by to ještě napínavější. Tedy jako kdybychom opravdu tiskli z původní výstupní sestavy.
- Nabídla se možnost zpracovávat TXT nebo XML, ale zde by problémy byly identické jako u HTML. Výhoda by byla, že by se mohlo začít příjemně TEXovat, ale hrajte si s něčím takovým, když na to máte několik hodin.
- Podobnou hořkost jsem ucítil u CSV či TXT, kde by se vše dalo importovat například do OpenOffice.org Calcu/Base či Microsoft Excelu/Accessu. Nastavit variabilní záhlaví je spíše o VBA než o přímé práci v buňkách.
- Přímá možnost TEXování se všemi příchutěmi parsování a umísťování objektů mě tentokrát také přešla poměrně rychle. Čas běží v takových momentech neskutečně rychle!

Bádání nad PDF

Budiž to PDF! Jenže ani Adobe Acrobat 8 Professional neumí nahrazování textového řetězce za jiný. Neumí to ani Adobe Acrobat 9 Pro Extended, který jsem zkusil v rámci třicetidenní zkušební doby. Pokud to umí, tak jsem na to v čase mně daném nepřišel. To mě poprvé zamrazilo. Ale i tak jsem si říkal, že zůstanu u PDF, že je to zajímavý problém a možná to někdo vyřešil. Vždyť těch e-knih je všude kolem dost a dost a je velká pravděpodobnost, že to již někdo potřeboval.

První krůčky vedly k programu pstoedit, `http://www.pstoedit.net/`. Ten umí velmi zdařile exporty z mnoha a do mnoha formátů. S autorem jsem dříve kratičce komunikoval, a tak jsem si říkal, že kdybych se někde zasekl, že by mi možná poradil. Bohužel po několika pokusech exportů do METAPOST, GnuPlotu, SVG, PS či jiných formátů, vše vedlo k nekonečnému rozhodovacímu stromu možností, jak v editacích a kompletacích dál.

Jak to tak v životě bývá, člověk v nouzi začne googlit. Experimenty mě zavedly k programu PDFill PDF Editor, `http://www.pdfill.com/`. Je to příjemná utilitka jen pro Microsoft Windows (borce od Linuxu odkazuji na emulátor Wine). Věděl jsem, co hledám. Slovíčko „Replace“, jak to zná na nahrazení řada editorů a DTP programů. Smůla! Nic takového k nalezení v tomto programu není. Výhoda byla, že šlo snáze vykreslit bílý obdélník, kdyby na to přišlo, ještě rychleji než v komerčním Adobe Acrobatu neboť vlastnosti vyskakují automaticky a v jedné kartě. Nouzová varianta zajištěna. Tento program umí řadu editačních operací, ale bez možnosti zasáhnout do textů přímo. Můžeme si to představit tak, že naše PDF je vrstva, na kterou kreslíme vlastní vrstvu. Srdíčko však zahřál seznam možností, které lze otevřít přes *Tools–PDF Tools (No Watermark)*.

V této souvislosti člověka napadá použít balíček `pdfpages` a volbu `pagecommand` nebo `picturecommand` s vykreslováním bílých obdélníků, nebo například jen silných bílých čar. Použití cyklu vítané. Problém tu však byl. Nad a pod nahrazovanými texty byly linky a další texty, do kterých se nesmělo zasáhnout. To by nebyl problém zabělit obdélníky, kdyby se texty mezi stránkami nehýbaly. Poněvadž se výška záhlaví tabulek měnila, nebyli jsme schopni tuto metodu použít bez toho, abychom dělali výškovou korekci na každé stránce. A nejen výšku, ale i počet bělících obdélníků na každé jednotlivé stránce.

Cvičení: Zkuste si tento přístup. Nakreslete si cvičně několik bílých obdélníků do PDF stránky načítané balíčkem `pdfpages`.

Mé další kroky vedly k programu nazvanému Free PDF Editor, `http://www.freepdfeditor.net/`, který však není editorem v pravém slova smyslu. Hezký a výstižný název programu, ale na tento druh úkolu byl nepoužitelný.

Podobně zkolaboval také nástroj pdftk, `http://www.accesspdf.com/pdftk/`, který sice umí práci dávkově, ale ne na úkony, které byly potřeba.

První dílčí úspěch zaznamenal až editor Foxit PDF Editor, `http://www.foxitsoftware.com/pdf/editor/`. Zkušební verze programu mi dokázala, že mohu editovat jednotlivé texty jako objekty a navíc mohu texty měnit. Tohle byl výrazný úspěch. Jistota žádného bělení; jen označení a mazání objektů. Přes klávesu Ctrl a Shift jde vícenásobné označení a odznačení. Také jde hromadné označení objektů, jak to známe z grafických programů. Mazání? Ano, jistě, klávesou Delete.

Ačkoliv je program komerční, lze jej třicet dní testovat. Začínal jsem jásat, i kdybych úkol nevyřešil civilizovaně, tak jsem se naučil něco nového. Navíc volba čištění bílými obdélníky byla otevřená i zde, přes *Object–Add Graphics–Add Filled Rectangle*.

Úspěchy Čechoslováků

Během testů mě osvítila jiná možnost. Program PDFedit, `http://pdfedit.petricek.net`, který cituje na stránkách ConTeXtu samotný Luigi Scarso, viz `http://wiki.contextgarden.net/User:Luigi.scarso#Luatex_examples`.

Bohužel můj CygWin se se mnou odmítal bavit a já neměl sílu si grafický režim rozhodit testy nad Qt3. Na druhém, linuxovém stroji, kde je něco, co pamatují historici, tedy Mandrake (nyní Mandriva), jsem časově zkolaboval na nutnosti kompilovat Céčkové knihovny. To jsem věděl, že není v mých časových možnostech. Název knihovny `replace_text-tool.exe` mi však zlepšil náladu.

Po tomto velkém dnu jsem na Mandrivě 2009.1 PDFedit rozjel bez nejmenších problémů užitím standardního rpmdrake. Ikonku lze nalézt pod *Grafika–Více–PDFedit*. Až poté jsem testy a četbou manuálu zjistil, že program umí editovat objekty, ale hromadné nahrazování textů zatím neumí. Umí však jejich úpravu. Práce v dávkovém režimu možná; slibný to nástroj.

Jak mi později potvrdil jeden z autorů a vývojářů PDFedit, Michal Hocko, tak na grafické úrovni to jejich program neumí, každopádně funkcionalita v jedné z knihoven obsažena je. Je to v knihovně, která je součástí zdrojového balíku. Ukázka, jak lze nahrazování docílit, je zmíněna v jednom z příkladů.

Cvičení: Během experimentů jsem zjistil, že PDFedit neumí otevřít PDF vzniklé za pomoci ConTeXt MkIV s aktivovaným LuaTeXem. To stojí samozřejmě za ověření a další bádání.

Blýská se na lepší časy

Rozsvítilo se u dalšího nástroje. Tím byla testovací verze německého programu PDF Editor 2.6, `http://www.cadkas.com/`, detaily viz [1]. Jedná se o demoverzi programu bez časového omezení. Tvůrci nejsou žádní začátečníci. Po instalaci jsem zjistil, že si mohu přepnout jazykovou mutaci *Sprache–English* (čeština je nabízena, ale není přeložená). Přepnul jsem se do angličtiny. To už jsem z hecu volal kolegovi, ať brousí němčinu, že jsem našel řešení v menu pod *Bearbeiten–Suchen und Ersetzen (in allen Textobjekten)*.

Vstupní PDF soubory jsem otevřel ze záložky *Open PDF file*. Program totiž nabízí nahrazování textu za text a relativně to funguje. Problém byl v tom případě, kdy jsem v tiskových možnostech *Předvolby tisku...–Upřesnit...–Písmo TrueType:* měl nastavené *Substituce písmem zařízení*. Pokud jsem si nastavil *Načíst jako písmo*, program dokázal vyhledávat textové řetězce obsahující znaky s diakritikou. Tohle byl základ. Nevýhodou zůstalo, že v uloženém PDF se objevily pruhy „Verändert mit der DEMOVERSION von CAD-KAS PDF-Editor (http://www.cadkas.de).“ To rovnou třikrát, rovnoměrně ve čtvrtinách strany bez horní části.

Tyto bonusové texty by šly vymazat jiným editačním nástrojem, který by mohl mít zase své texty nebo loga atd. Na tohle jsem však rychle zapomněl, neb mě tehdy zatlačil čas a cokoliv dělaného ručně bylo nepřípustné.

Cvičení: Program je chráněn autorskými právy, nesmí být dále šířen a prodáván. V této chvíli by to chtělo právníka, jestli lze zasahovat do výstupů z takového zkušebního programu. Jinými slovy, zdali můžeme pokračovat v editacích takto uloženého souboru. Mé znalosti říkají, že ano, dokud to licence nezakazuje. Náš pan právník mi to objasnil slovy, že demoverze programů slouží právě na to, aby si je uživatel mohl vyzkoušet před vlastní koupí. Například výstupy z programu SPSS jsou součástí programu, což není tento případ, zde dodávám své vstupní PDF k editaci. V mém případě byly výstupy použity k tisku a PDF soubory jako takové nebyly zachovány; zde to nehrálo roli.

Confessio est non probatio ergo ab instanitia absolutio. Ad acta! (Přiznání není důkaz tedy osvobozen pro nedostatek důkazů. Považujme za vyřízené!)

Samozřejmě přidání takových reklamních textů (někdy i obrázků) programem automaticky znemožňuje jejich okamžitý tisk a zajištění identity. To dělá ostatně řada tiskových ovladačů a demoverzí programů.

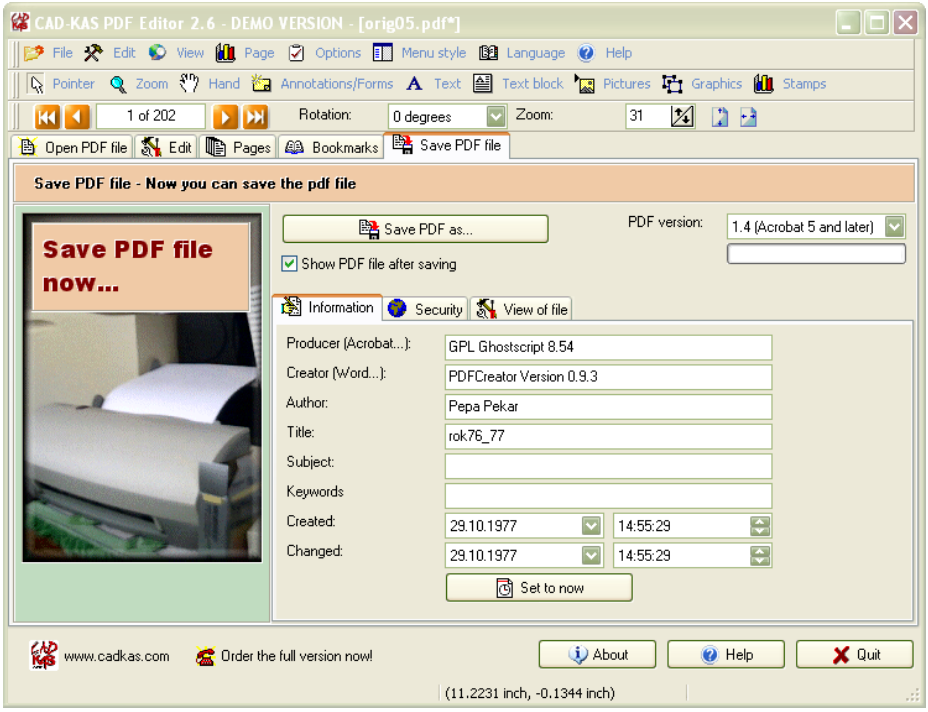
V tomto případě jsem použil následující fintu s využitím výhody škálovatelnosti vektorové formy. Před nahrazením textů jsem vstupní PDF umístil na A1 formát a A4 stranu jsem posunul do oblasti, kde se informační texty neobjeví. Poté stačí texty nahradit, uložit výsledek a PDF oříznout.

První krok umístění na A1 a posun se dá zrealizovat balíčkem pdfpages. Problémem ale je, že tento balíček tiskovou stranu obalí jako objekt a program PDF Editor není schopen zasáhnout do vlastních textů. Zde jsem použil demoverzi nebo komerční Adobe Acrobat 8 Professional, tedy z menu vybrat *Dokument–Oříznout stránky...* V bloku *Změnit velikost stránky* jsem nastavil velikost stránky A1. Odškrtl jsem možnost *Na střed* a nastavil *Posun X* na nula palců (0") a *Posun Y* na 20 palců (20"). Soubor jsem uložil.

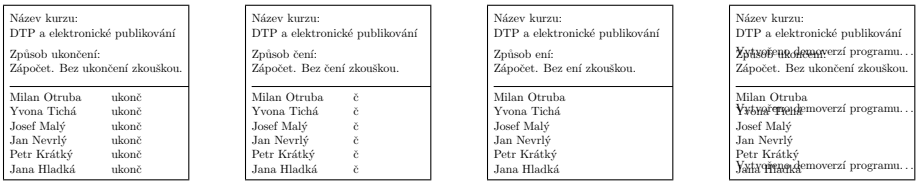
Tento soubor jsem načetl v PDF Editoru. Zvolil jsem z menu *Edit* a poté *Search and Replace (within each text object)*. Do *Search text* a *Replace text* jsem si nastavil co a za co změnit.

Na ukázku zmiňme smazání řetězce „Ukonč“. Ne však dalších výrazů (Bez ukončení, Způsob ukončení, Ukončený, Ukončit, Neukončit atd.). Lze to obejít dočasným nahrazením. Například „Bez ukončení“ nahradit „pivopivopivo“; „Způsob ukončení“ nahradit „vinovinovino“. Jsou to textové řetězce, o kterých jsme si jistí, že se v dokumentu nevyskytují. Poté se nahradí „Ukonč“ za prázdný řetězec a zpětně se nahradí „pivopivopivo“ na „Bez ukončení“ a na závěr „vinovinovino“ za „Způsob ukončení“.

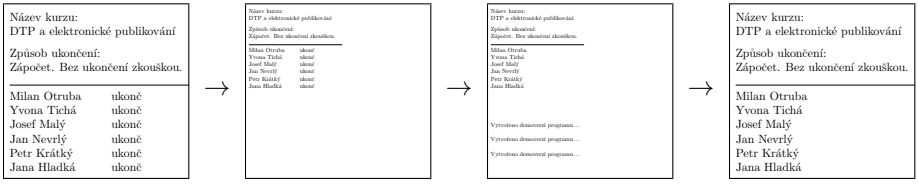
Jak vidíme, nahrazujeme jednoduchý řetězec a tolik je s tím problémů. Proto by se hodila práce s regulárními výrazy a dávkově, ale tento program to neumí. To zjistíme snadno zapsáním do příkazového řádku `PDFEdit.exe --help`, kdy



Obrázek 1: Demoverze programu PDF Editor 2.6 v akci.



Obrázek 2: Originál, úprava o „ukon“, „ukonč“ a zdařilá úprava s texty navíc.



Obrázek 3: Originál, zmenšení, násobné úpravy a výsledný ořez.

se otevře grafické okno. Velkou naději dává právě program PDFedit od českých a slovenských tvůrců.

Cvičení: Za zajímavý úkol pro svobodné programy stojí spojování PDF tak, aby objekty nebyly zabaleny. To samé platí na zmenšení velikosti či posunutí stránky. Tipem nechť je program pdftk, `http://www.accesspdf.com/pdftk/`.

Cvičení: Zkuste změnit metriku PDF stran v programu PDFedit a zjistěte možnost editace v programu PDF Editor.

Stihl jsem ještě ověřit starší verze tohoto programu, verzi 2.4 a mobilní verzi 2.6. Bohužel v jednom případě je výstupní PDF z demoverze programu rastrováno a se ztrátovou kompresí uloženo; v novější verzi nelze vyhledat znaky s diakritikou, proto slovo „Ukonč“ nelze touto cestou vůbec nalézt. Nahrazení jen části „Ukon“ by nám nepomohlo, poněvadž znak „č“ by v PDF souboru zůstal a muselo by se do dalších (ručních) zásahů.

Demoverze programu verze 2.6 má drobné mouchy. V jednom PDF souboru mi to množilo dvojtečky. Takže po pěti konverzích jsem z „:“ měl „::::::“, to však šlo jednorázovým nahrazením „::::::“ na „:“ zdařile vyčistit. Podobně to platilo na „::“, „:::“ atd.

Uložení je možné ze záložky *Save PDF file*, kde bylo také možné změnit metadata PDF souboru, včetně data vytvoření a změny souboru, viz první obrázek na straně 20. Nu vida, ani změna systémového času by nebyla třeba.

Závěrečný krok bylo zpětně ořezat A1 formát (s jednou A4 stranou blízko levého horního rohu) na tisknutelné A4. Lze použít opět Adobe nebo zde se náramně hodí balíček pdfpages, kdy nám již nevadí obalování PDF stránek před vlastním tiskem. Já si na závěr denní práce vychutnal samozřejmě variantu nekomerčním produktem.

```
% pdflatex vystup.tex
\documentclass[a4paper]{article}
\usepackage{pdfpages}
\begin{document}
\includepdf[pages={-},viewport=0 1518 582 2218]{vstup.pdf}
\end{document}
```

Demonstraci celého problému na smyšlených datech vysokoškolského kurzu a kroky jeho řešení lze nalézt na druhém a třetím obrázku na straně 20.

Příběh s dobrým koncem

Upravené PDF jsem kolegovi večer předal a měl to ráno připravené na tisk do svého archivu.

Tato pohádka má dobrý konec, ale poučením nechť nám je, že nemáme hned všem všechno slibovat. Také je vidět, že v časové nouzi člověk používá absolutně

všechno, co má při ruce, a to včetně legálně koupených komerčních nástrojů, když zrovna neví, po kterém Open Source Software sáhnout.

Také je výhodné studium nástrojů a testy na vlastních cvičných problémových situacích. Realita není pohádka a procházka růžovým sadem, proto na papíře a v archivech vypadá vše tak krásně. Třeba tisky na laserové tiskárně v archivech zpětně uložené a datované do doby, kdy laserové tiskárny ještě neexistovaly!

Reference

[1] Domovská stránka programu PDF Editor 2.6. [online, cit. 6. června 2010] Dostupné z `http://www.cadkas.com/pdf-editor-edit-pdfs.php`, demoverze ke stažení z `http://www.cadkas.com/pdfedit!.exe`.

Summary: Editing PDF File

The real-world problem of deleting specific text parts in PDF files of hundreds of pages occurred out of the blue sky and the deadline was to finish the task within 24 hours. This article presents our experience with editing those PDF files using different proprietary software and trial versions as well as tools and programs from the world of open source software.

Keywords: Text replacing, PDF editing, Adobe Acrobat, PDFedit, CAD-KAS PDF Editor 2.6, pdfpages package.

Pavel Stříž, striz@fame.utb.cz
Radek Benda, benda@fame.utb.cz
ÚSKM FaME UTB ve Zlíně, Mostní 5139
Zlín, CZ-760 01, Czech Republic