Portál pre odborné publikovanie ISSN 1338-0087

Možnosti konverzie súborov typu TEX využívajúcich slovenský jazyk do formátov založených na XML

Zahradníková Barbora · Informačné technológie

14.06.2013

Pôvodné kódovanie LaTeX-u bolo vyvinuté v 70. a 80. rokoch pre potreby textov v americkej angličtine, a preto nezodpovedá mnohým požiadavkám slovenského jazyka. Splnenie týchto požiadoviek možno zabezpečiť správnou voľbou kódovania doplnenou zodpovedajúcimi fontami. Keďže bol LaTex primárne vyvinutý pre písanie textov v anglickom jazyku, programy zaoberajúce sa jeho konverziou taktiež nie sú prispôsobené konverzii súborov využívajúcich iné ako defaultné kódovania.

1. TeX4ht

Jedným z najpoužívanejších programov pre konverziu súborov písaných vo for- máte typu TeX do formátov založených na XML formáte je TeX4ht. Bol vyvinutý na Ohio State University a umožňuje konverziu do formátov ako HTML, DOCBOOK alebo ODF. Jednotlivé submoduly programu možno použiť buď volaním celej cesty k súboru (napr.: usr share tex4ht oolatex filename.tex) alebo využitím skriptu mk4ht (napr.: mk4ht oolatex filename.tex). Pri využití programu TeX4ht na konverziu súborov vykazujúcich typické črty slovenčiny, príp. češtiny dochádza k viacerým chybám súvisiacim so zvoleným kódovaním a s nekompatibilitou niektorých balíkov s programom.

Druhy kódovaní v LaTex-u

Pre správne zobrazenie súborov písaných v LaTex-u je potrebné zosynchronizo- vať 3 druhy kódovaní:

1. inputencoding–vstupné kódovanie, ktoré je špecifikované používateľom vo vytvorenom súbore, zvyčajne príkazom inputenc. Toto kódovanie zabezpečuje, že znaky zadané na klávesnici program rozpozná a môže ich ďalej spracovať
2. internal encoding – je vnútorné kódovanie LaTeX–u, čiže kódovanie, ktoré je aplikované pri samotnej kompilácii LaTeX–om. Predstavuje reprezentáciu znakov v LaTeX – u
3. font encoding – kódovanie fontov, predstavuje priradenie znakového kódu ku glyfu (vonkajšia reprezentácia znaku), zabezpečuje zobrazenie správneho znaku pri sádzaní výstupného textu podľa zvoleného fontu.

3. Možnosti implementácie slovenčiny

Vstupné kódovanie je zvyčajne zhodné s kódovaním operačného systému. Pre OS

Windows je typické kódovanie CP-1250, pre unixové systémy je to ISO/IEC 8859-2. V snahe eliminovať problémy, ku ktorým dochádza kvôli nekompatibilite rôznych kódovaní, prišlo k snahe vytvoriť kódovanie, ktoré je nezávislé od použitého operačného systému. Odporúčaným kódovaním nielen pre slovenčinu ale pre všetky jazyky vrátane gréckeho, ruského či čínskeho jazyka sa preto stáva UTF-8. Toto kódovanie predstavuje univerzálne riešenie nekompatibility kódovaní rôznych jazykov nakoľko nielenže umožňuje využívanie znakov rôznych jazykov, ale zároveň ich kombinovanie. Preklad vstupného kódovania do vnútorného jazyka LaTeX – u je zabezpečený balíčkom inputenc.

```
 usepackage[encoding name]{inputenc}
```

3.2 Výstupné kódovanie

Pre zabezpečenie správneho výstupného kódovania a následného zobrazenia znakov je potrebné zadeklarovať rodinu fontov a spôsob kódovania fontov. LaTeX2ε prekladá znaky z vnútorného kódu do výstupného kódovania na základe v súbore zadeklarovaných prekladových tabuliek:

```
 usepackage {rodina fontov}
 usepackage [kódovanie fontov]{fontenc}
```

Pre správnu implementáciu slovenského jazyka sú dve možnosti deklarácie výstupného kódovania:

Babel

Prvou z možností je implementácia balíčka babel. Jeho snahou je systémovo vyriešiť problém nekompatibility jazykov a umožniť používateľovi nielen písať v nejakom konkrétnom jazyku, ale tiež vo viacerých jazykoch naraz bez výrazných problémov. [30] Balíček Babel obsahuje algoritmy delenia slov a odsadzovania odstavcov v použitých jazykoch, preklady au- tomaticky generovaných textových elementov (napr. nadpisy Obsah, Literatúra, Kapitola, atď.) ako aj formáty pre správnu tlač dátumu a časových údajov. Odporúčaná hlavička pri použití slovenčiny resp. češtiny v babelizovanom LaTeX–u je:

```
documentclass{article}
usepackage[utf8]{inputenc}
usepackage[slovak]{babel}      %      alebo      
usepackage[czech]{babel} usepackage[T1]{fontenc}
usepackage[názov vstupného kódovania]{inputenc}
... %zvyšok hlavičky
begin{document}
...
end{document}
```

Konverziu súborov využívajúcich balíček babel prostredníctvom programu Tex4ht zabezpečujú TeX4ht submoduly, ktorých názov končí názvom kompilátora “latex”, čiže napr. oolatex, htlatex, dblatex, xhlatex,... Pri konverzii prostredníctvom programu TeX4ht možno pri použití balíka cite pozorovať nefunkčnosť prepojenia odkazov pri

citáciách so zodpovedajúcimi bibliografickými záznamami. Pri plánovanej TeX4ht konverzii súborov ob- sahujúcich balík babel je potrebné pre správne prepojenie odkazov s cieľmi implementovať namiesto balíka cite balíček hyperref. Ich použitie je podobné, preto vo väčšine prípadov postačuje zakomentovať v hlavičke súboru riadok

```
usepackage{cite}
```

a nahradiť ho

```
usepackage{hyperref}
```

Pri definovaní slovenského jazyka prostredníctvom babel–u dochádza tiež k nerozpoznaniu jednoduchých (anglických) úvodzoviek, resp. apostrofu (’) a znaku ” ˆ ” . Z tohto dôvodu nie sú rozpoznané ani niektoré preddefinované skratky (shorthands). Pokiaľ používateľ nemá nainštalovanú slovenskú, resp. českú klávesnicu a chce použiť znak s diakritikou (napr. “ó”), nestačí zadať skrátený tvar “’o”, ale musí použiť štandardné príkazy “{’}o” alebo {textquoteright}o.

csLaTex

CsLaTeX je sada konfiguračných súborov pre LaTeX, ktoré umožňujú pripravovať české a slovenské dokumenty. Zabezpečuje konverziu z vstupného kódovania na vnútorné kódovanie ISO 8859-2, použitie správnych fontov (csfontov), zároveň obsahuje vzory pre delenie anglického, českého a slovenského jazyka, zaisťuje tiež nastavenie rovnomerných medzier, nastavenie slovenských názvov sekcií či spávne nastavenie dátumu. Hlavička dokumentu používajúceho csLaTeX môže vyzerať:

```
documentclass{article}
usepackage[utf8]{inputenc}
usepackage{slovak} % alebo usepackage{czech}
usepackage[T1]{fontenc}
... %zvyšok hlavičky
begin{document}
...
end{document}
```

Pre správne fungovanie je potrebné, aby bol balíček {slovak} zadeklarovaný skôr ako rodina fontov a spôsob kódovania fontov. TeX4ht konverzia súborov využívajúcich balíky slovak, resp. czech nie je súčasťou distribúcie, preto bolo potrebné submoduly pre csLaTeX doplnit do zdrojových kódov. Spolu s návodom na inštaláciu je možné ich stiahnuť na konci článku. Použitie je podobné ako pri ostatných submoduloch, čiže buď volaním prostredníctvom celej cesty alebo skriptom mk4ht. Rozdiel spočíva len v názve volaného submodulu. Názvy submodulov využívajúcich kompilátor csLaTeX končia reťazcom “cslatex”.

Príklad: Pre konverziu do ODF, OPEN DOCUMENT FORMAT, stačí v príkazovom riadku zadať:

```
mk4ht oocslatex filename.tex
```

4. Ďalšie poznámky ku konverzii prostredníctvom TeX4ht

Použitie usepackage{lmodern} pri plánovanej konverzii prostredníctvom oolatex-u (resp. oocslatex-u) nie je vhodné bez ohľadu na zvolené kódovanie. Výsledkom takejto kompilácie je chyba (— OoFilter Error 6 — Improper record: Font("lmsy8″,"",”8″,”100″)

Pri súčasnej implementácii balíka wrapfig a konverzii do ODF formátu (čiže súboru typu *.odt) dochádza k chybe kvôli snahe balíka wrapfig obtekať matematické prostredia deklarované v súbore. Pre elimináciu tejto chyby je potrebné matematické prostredia definovať vovnútri iných blokov, inak je vypísaná chyba(— xtpipes error 29 — At <sax content-handler= “xtpipes.util.ScriptsManager, tex4ht.OoFilter” lexical-handler=”xtpipes.util.ScriptsManagerLH”>). Príklad správnej implementácie matematických prostredí:

```
begin{center}
begin{equation}
...
end{equation}
end{center}
```

alebo

```
begin{center}
begin{math}
...
end{math}
end{center}
```

Literatúra

1. MITTLEBACH, F.:LaTEX2ε Encoding Interfaces. Purpose, Concepts, and Open Problems. 1995. p. 3-17. [cit. 1.5.2013]. Dostupné na internete: ftp://ftp.tex.ac.uk/pub/tex/macros/latex/doc/encguide.pdf
2. LaTeX2e font selection. LaTeX3 Project Team. [cit. 3.5.2013]. Dostupné na internete: http://tex.loria.fr/general/new/fntguide.html
3. Olšák, P.: Manuál k CSTEXu. 2012. p. 38-47. [cit. 6.1.2013]. Dostupné na internete: math.feld.cvut.cz/olsak/ftp/cstex/doc/cstexman.pdf
4. Braams, J.:Babel, a multilingual package for use with LATEX’s standard document classes. 2006. Zoetermeer. p. 1.[cit. 6.1.2013]. Dostupné na inter- nete: parokia.kre.hu/lelkesz/latex/babel.pd
5. Babel vs. CSLaTeX. home.ef.jcu.cz. [cit. 6.1.2013]. Dostupné na internete: http://home.ef.jcu.cz/~houda/miktex/install/babel-cslatex. html

---

Spoluautorm článku je Peter Fodrek