Dokumentace úlohy CSV: CSV2XML v Perlu do IPP 2011/2012
Jméno a příjmení: Peter Lacko
Login: xlacko06

Na zpracování argumentů příkazového řádku je použit standardní modul `Getopt::Long`, jelikož ale neumožňuje rozpoznat stejné parametry, je nutno provést tuto kontrolu manuálně. Argumenty jsou za použití modulu `Encode` překódované do znakové sady UTF-8. Následně se program pokusí o otevření vstupního a výstupního proudu, které se v případě úspěchu upraví funkcí `binmode()` tak, aby umožňovaly číst/zapisovat data v kódování UTF-8. Následuje kontrola validity hodnot zadaných parametrů (kořenový element, řádkový element, separátor, startovací index), nahrazení nekorektních znaků v hodnotě parametru `--missing-value`. Všechny testy jsou prováděny pomocí regulárních výrazů.

Informáce o činnostech vykonaných v případě chybného vstupního souboru, se ukládají do proměnné `$recovery`, defaultně nastavené na nulu. Pokud byl zadaný parametr `-e`,`--error-recovery`, proměnná `$recovery` se na základě dalších parametrů nastaví následovně:

- `--missing-value`: `$recovery = $recovery OR 1`
- `--all-columns`: `$recovery = $recovery OR 2`
- `-h`: `$recovery = $recovery OR 4`

Pokud nebyl zadaný parametr `-n`, do výstupního souboru se zapíše hlavička generovaného xml souboru, jestliže naopak byl zadán parametr `-r`,`--root-element` do souboru se zapíše otevírací tag kořenového elementu (podobně je to i s uzavíracím tagem).

Na zpracování csv souboru je využit modul `Text::CSV`, který má na starosti ověření správnosti každého řádku csv souboru a jeho rozdělení na sloupce. Vzhledem k tomu že se modul neřídí striktně předepsaným formátem souboru[1] a povoluje jako řádkové zlomy i znaky LF místo CRLF, a znak nového řádku na konci souboru, byla vytvořená proměnná `$crlf` indikující znak načtení korektního konce řádku. První řádek souboru se zpracovává samostatně, jelikož určuje počet sloupců a v případě zadaní parametru `-h` také názvy jednotlivých sloupcových elementů. Bez zadaného `-h` jsou tyto názvy generovány jako `colX` (kde X=(1 nebo index_zadany_parametrem) ... pocet_sloupcu) a uloženy do pole. Zpracování rozparsovaných řádků má na starosti funkce `$handle_row()`, která je v cyklu volána na každý řádek zvlášť. Jako parametry se funkci prodávají:

- odkaz na pole názvů sloupcových elementů,
- odkaz na pole získané rozparsováním řádku,
- proměnná `$recovery` popsaná výše a
- náhradní hodnota elementu pro chybějící sloupce zadaná parametrem `--missing-value`

Funkce nahradí problémové znaky `'&'`, `'<'`, `'>'`, `'"'` řetězci `"&"`, `"<"`, `">"` a `"&quot"`. Chybějící hodnoty sloupcových elementů funkce doplní hodnotami předanými čtvrtým parametrem, přebytečná odstrání nebo obalí vhodnými párovými elementy. Na korektní ukončení programu slouží funkce `ExitProgram()`, která uzavře otevřené soubory a vráti hodnotu jí předanou parametrem jako návratovou hodnotu programu.

Program byl kvůli čistotě a přehlednosti kódu psaný a laděný v režimu strict (příkaz `use strict;`) a se zapnutým výpisem varovných hlášení (příkaz `use warnings;`).

[1] Viz http://tools.ietf.org/html/rfc4180