Dokumentace úlohy CSV: CSV2XML v Perlu do IPP 2011/2012
Jméno a příjmení: Radek Ševčík
Login: xsevci44

1 Úvod

Tento dokument popisuje implementační řešení perlovského skriptu s podporou UTF-8, který slouží ke konverzi formátu dokumentů CSV do XML. Podoba výsledného XML dokumentu může být ovlivněna použitými přepínači. Je nutné vzít v úvahu nevalidní CSV dokumenty a také se vyhnout zakázaným znakům v párových značkách XML dokumentů.

2 Popis řešení

2.1 Parametry příkazové řádky

Zpracování parametrů je prvním krokem běhu programu, zajišťuje to modul `getopt`. V této fázi se kontroluje nejen správná kombinace zadaných parametrů, ale také jejich hodnoty. Neošetření vstupu by mohlo vést k rozhození struktury (zápisem nepovolených znaků) XML dokumentu.

Skript se chová jako textový filtr, tzn. není-li zadán vstupní/výstupní soubor parametrem, provede se čtení/zápis standartním vstupem/výstupem.

2.2 Národní prostředí

Chceme-li podporovat znaky národní abecedy, tak musíme načíst modul `locale` a POSIXovou knihovnu `locale_h`. Poté zavoláme v programu funkci `setlocale(LC_ALL, '')`, která si nastaví locales dle prostředí. Aby bylo možné používat kódování UTF-8, musíme pro parametry příkazové řádky a buňky CSV využít funkci `utf8::decode`.

2.3 Zpracování CSV dokumentu

Jako parser slouží modul `Text::CSV`. Je poměrně jednoduchý a nemá problém se zpracováním UTF-8 souborů (binární mód). Načítání probíhá po řádcích, které by měli být zakončené CRLF (mimo posledního), ale většina knihoven to striktně nedodržuje, proto se zachováme z důvodu kompatibility stejně.

První řádek určuje počet sloupců, jako oddělovač se implicitně použije čárka (nastavení v modulu), ale přepínačem se dá vynutit jiný znak. Tento řádek může také tvořit hlavičku, která se použije při generování XML. Pokud je v dalších řádcích méně či více sloupců, a je použitý přepínač pro zotavení se z chyb, pak tyto chybějící sloupce do XML doplníme určitou hodnotou (implicitně mezera) a přebývající ignorujeme, popř. je zahrneme do XML všechny.

2.4 Generování XML dokumentu

Jako XML writer byl vybrán `XML::LibXML`, používá přístup DOM (Document Object Model) a proto se s ním lehce manipuluje. Umí sám nahradit zakázané znaky např. `& < >`, ale i přes to dokáže vygenerovat nevalidní XML v názvech elementů. Proto se musí tyto vstupy ošetřit.

V závislosti na přepínačích se může vynechat hlavička či zadat název kořenového elementu, to je užitečné např. chceme-li výsledky kompletovat. Může se také nastavit název elementu obalující jednotlivé řádky nebo zaznamenat číslo řádku do atributu index (prvotní hodnota čítače se mění parametrem).

3 Závěr

Tento skript byl vyvíjen a otestován na operačním systému CentOS 5.8 x64 (GNU/Linux) sadou přiložených testů s nastaveným prostředím `export LC_ALL=cs_CZ.utf8`.

Reference

[1] Y. Shafranovich: *Common Format and MIME Type for CSV Files*, 2005. <http://www.ietf.org/rfc/rfc4180.txt>

[2] T. Bray, J. Paoli, E. Maler, F. Yergeau, C. M. Sperberg-McQueen: *Extensible Markup Language (XML) 1.0* Fifth Edition, 2008. <http://www.w3.org/TR/xml/>