Zadání úlohy do projektu z předmětu IPP 2014/2015

Zbyněk Křivka a Dušan Kolář
E-mail: {krivka, kolar}@fit.vutbr.cz, {54 114 1313, 54 114 1238}

CSV: CSV2XML

Zodpovědný cvičící: Zbyněk Křivka (krivka@fit.vutbr.cz)

1 Detailní zadání úlohy

Vytvořte skript pro konverzi formátu CSV (viz RFC 4180) do XML. Každému řádku CSV bude odpovídat jeden dodefinovaný párový element (viz parametr `-l`) a ten bude obsahovat elementy pro jednotlivé sloupce (viz parametr `-h`). Tyto elementy pak již budou obsahovat textovou hodnotu dané buňky z CSV zdroje. Pro XML problematické znaky s UTF-8 kódem menším jak 128 v obsahových buňkách CSV konvertujte na odpovídající zápisy v XML pomocí metaznaku `&` (např. `&`, `<`, `>` atd.). Ostatní problematické znaky konvertovat nemusíte.

Tento skript bude pracovat s těmito parametry:

- `--help` viz společné zadání všech úloh.
- `--input=filename` zadaný vstupní CSV soubor v UTF-8.
- `--output=filename` textový výstupní XML soubor s obsahem převedeným ze vstupního souboru.
- `-n` negenerovat XML hlavičku[1] na výstup skriptu (vhodné například v případě kombinování více výsledků).
- `-r=root-element` jméno párového kořenového elementu obalující výsledek. Pokud nebude zadán, tak se výsledky neobalují kořenovým elementem, ač to porušuje validitu XML (skript neskončí s chybou). Zadání řetězce `root-element` vedoucího na nevalidní XML značku ukončí skript s chybou a návratovým kódem `30`. Nevalidní znaky nijak nenahrazujte.
- `-s=separator` nastavení separátoru (jeden znak) buněk (resp. sloupců) na každém řádku vstupního CSV, kromě tisknutelných znaků jako například mezera, středník či čárka podporujte i identifikátor TAB pro tabulátor jako oddělovač; implicitní hodnota je znak čárka (`,`). Volba intuitivně rozšiřuje RFC 4180.
- `-h=subst` první řádek (přesněji první záznam) CSV souboru slouží jako hlavička a podle něj odvoďte jména elementů XML. Každý nepovolený znak nahraďte řetězcem *subst* (vznikne-li i tak invalidní XML element, skončete s chybou a návratovým kódem `31`). Je-li `-h` zadáno pouze jako volba, je *subst* implicitně pouze znak pomlčka (`-`). První záznam CSV se potom již neobjevuje jako obsah ve výsledném XML. V případě, že tento parametr/volba zcela chybí, budou jména elementů pro buňky (resp. sloupce) generovány dle parametru `-c=column-element`.

[1] Tradiční XML hlavička je `<?xml version="1.0" encoding="UTF-8"?>`

- `-c=column-element` určuje prefix jména elementu `column-elementX`, který bude obalovat nepojmenované buňky (resp. sloupce), kde X je inkrementální čitač od 1. Implicitní hodnotou `column-element` je řetězec `col`. Zadání řetězce `column-element` vedoucího na nevalidní XML značku ukončí skript s chybou a návratovým kódem `30`.

- `-l=line-element` jméno elementu, který obaluje zvlášť každý řádek vstupního CSV; implicitní hodnota je `row`. Zadání řetězce `line-element` vedoucího na nevalidní XML značku ukončí skript s chybou a návratovým kódem `30`. Nevalidní znaky tentokráte nenahrazujte.

- `-i` zajistí vložení atributu `index` s číselnou hodnotou do elementu `line-element` (tento parametr se musí kombinovat s parametrem `-l`; jinak nastane chyba kombinace parametrů).

- `--start=n` inicializace inkrementálního čitače pro parametr `-i` na zadané kladné celé číslo `n` včetně nuly (implicitně `n` = 1). Není-li tento parametr kombinován s `-i` a `-l`, ukončete skript s chybou parametrů.

- `-e`, `--error-recovery` zotavení z chybného počtu sloupců na neprvním řádku (první řádek bude sloužit pro odvození správného počtu sloupců) tj. každý chybějící sloupec bude doplněn prázdným polem, přebývající sloupce budou ignorovány. Pokud nebyl zadán tento parametr a vstup obsahuje na některém řádku špatný počet sloupců, tak skript ukončete s chybou a návratovým kódem `32`.

- `--missing-field=val` Parametr je povolen pouze v kombinaci s `--error-recovery` (resp. `-e`). Pokud nějaká vstupní buňka (sloupec) chybí, tak je doplněna zde uvedená hodnota `val` místo pouze prázdného pole. Problematické znaky konvertujte na odpovídající zápisy v XML pomocí metaznaků.

- `--all-columns` Parametr je povolen pouze v kombinaci s `--error-recovery` (resp. `-e`). Sloupce, které jsou v nekorektním CSV navíc, nejsou ignorovány, ale jsou také vloženy do výsledného XML. Pokud je tento parametr navíc v kombinaci s `-h`, tak sloupce, ke kterým nebyla prvním řádkem definována hlavička, budeme dle parametru `-c=column-element` značit `column-elementX`, kde X je pořadí sloupce na daném řádku (např. třetí sloupec, který byl na daném řádku první bez určené hlavičky, bude označen `column-element3`).

Verze formátu CSV, ač není standardizován, bude uvažována z RFC 4180 (sekce 2)[2].

Výtah a upřesnění RFC 4180: Záznamy (anglicky *record*) jsou odděleny koncem řádku (dvojznak CRLF, ne pouze LF). Poslední řádek je bez ukončení pomocí CRLF. Na prvním řádku mohou být volitelně popisky sloupců. Pokud to okolnosti vyžadují, tak se řetězce píší v uvozovkách (např. pokud řetězec obsahuje uvozovky, separátor buněk nebo separátor záznamů). Uvozovky v ouvozovkovaném řetězci se píší jako dva znaky uvozovek vedle sebe. Pokud buňka (anglicky *field*) obsahuje ouvozovkovaný řetězec, tak nesmí obsahovat žádné znaky mezi separátory a ohraničujícími uvozovkami[3]. Bílé znaky se nevynechávají[4].

Pro účely této úlohy je ještě nutné rozšířit definici neterminálu TEXTDATA z RFC 4180, aby akceptovala i UTF-8 znaky s kódem větším jak 127.

[2]Nekolizní výrazy „MAY/SHOULD BE“ je třeba v naší úloze chápat jako „MUST/HAVE TO BE,“ pokud to neupřesňují zadané parametry skriptu.

[3]Hodně knihoven pro zpracování CSV není takto restriktivních, takže toto pravidlo bude testováno pouze okrajově.

[4]Výjimka: V textových elementech jsou okrajové bílé znaky ignorovány nástrojem pro porování XML souborů.

Reference:

- Y. Shafranovich: RFC 4180 - Common Format and MIME Type for Comma-Separated Values (CSV) Files, 2005. Dostupné na `http://tools.ietf.org/html/rfc4180` [citováno 16. 2. 2010]

2 Bonusová rozšíření

- **PAD** (0,5 bodu): Podpora parametru `--padding`, který u použitých čitačů (řádků při kombinaci s `-i` a sloupců tvaru `colX`) provede doplnění takového počtu nul zleva, aby všechny čísla dané sekvence měla stejný počet číslic, který bude ale minimální dostačující. Sekvence pro číslování sloupců je pro každý záznam samostatná.
- **VLC** (1,0 bodu): Pokročilá validace vstupního CSV souboru vůči striktnímu výkladu RFC 4180 při zadání parametru `--validate`, který nelze kombinovat s žádným dalším parametrem kromě `--input` a `--output`. Pokud bude vstup invalidní, vracejte návratový kód `39`, jinak vracejte návratový kód `0` a na výstup vypište výsledek převodu validního CSV do XML.

3 Specifické požadavky na dokumentaci

Pokud použijete externí knihovnu, popište její nastavení, aby její činnost reflektovala zadání.

4 Poznámky k hodnocení

Výsledný XML soubor bude porovnáván s referenčními XML soubory nástrojem JExamXML na porovnání XML souborů, který se umí správně vypořádat například s různým odsazením elementů. Více viz stránka *IPP:ProjectNotes* na Wiki předmětu.